BASEMENT

MASSACHUSETTS INSTITUTE of TECHNOLOGY
LIBRARIES

HD28
.M414

No1303-82

Dewey

MASS. INST TECH.
JUL 29 1982
LIBRARIES

WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

# The Nature of Residential Telephone Usage and Reliability of Reporting

by

John H. Roberts

W.P. #1303-82 February 1982

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

# The Nature of Residential Telephone Usage and Reliability of Reporting

by

John H. Roberts

W.P. #1303-82 February 1982

M.I.T. LIBRARIES
JUL 2 9 1982
RECEIVED

Abstract

This paper examines long distance toll calling rates for residential telephone subscribers in Quad Cities, Illinois/Iowa. Accurate information on calling rates allows an examination of the degree to which toll calls follow certain distributional assumptions commonly made in marketing research concerned with frequently purchased goods and services. A number of shortcomings are found in the fit of the Poisson distribution to the number of calls in a period and the Normal distribution has a severely truncated tail at zero.

These distributional assumptions underlie models which relate perceived calling rates, reported by household heads, to actual usage. After the deletion of extreme observations the optimal estimate of a respondent's calls weights the survey response at about twice the level of the population mean.

Calling rates are not significantly related to socioeconomic group, but the reliability of reporting them is. Wives prove to be more reliable respondents than husbands in most segments. Differences between the responses of husbands and wives does not provide a good indication of accuracy but consistency between the average bill reported and the average calls reported do.

C744336

C744336

Table of Contents

1. Introduction

1.1. Objectives and Motivation
1.2. Model of Reported Calls on Actual Calls
1.3. Background and Data Collection

2. The Nature of Telephone Toll Call Usage

2.1. The Importance of Studying the Process
2.2. Interpurchase Times
2.2.1. First order Autocorrelation
2.2.2. Relationship Between Mean and Standard Deviation
2.2.3. Chi-Square Goodness of Fit Tests
2.2.4. Distribution of Errors
2.3. The Number of Calls in an Interval
2.3.1. Independence of Time Periods
2.3.2. Relationship Between Mean and Variance
2.3.3. The Gamma Assumption
2.3.4. The Normality Assumption
2.4. Conclusions

3. Reliability of Reporting

3.1. Bias and Distribution of Error
3.1.1. Relation of Actual Mean to Reported
3.1.2. A Variance Decomposition
3.1.3. Socioeconomic Differences
3.1.4. Accuracy of Reporting for Household Members
3.2. Husband-Wife Reliability
3.2.1. Relation of Actual Mean to Reported by Spouse
3.2.2. Socioeconomic Differences
3.2.3. Accuracy of Reporting for Household Members
3.2.4. Husband-Wife Differences as Measures of Reliability

4. Validity of Results

4.1. Internal Validity
4.1.1. Selection Biases
4.2. External Validity

5. Conclusion

5.1. Summary
5.2. Areas for Future Research
5.2.1. Compound Estimation Rules
5.2.2. Alternative Measures of Association
5.2.3. Modeling of the Billing Process
5.2.4. Reliability Measurement
5.2.5. Modeling of Homogenous Segments
5.2.6. Validity Considerations

Acknowledgements

Dean Alvin Silk was responsible for the concept of this project. He provided not only a suggestion for the general approach, but also helpful comments and a rich source of references as the work progressed. To him I am most grateful.

Jim Lattin was responsible for setting up the data base; a painful and arduous task. He spent many hours making the analysis possible and, having done so, was still prepared to spend many more discussing results and reading drafts.

The two readers, Professors Tony Wong and Ernie Berndt provided useful external insights from both a statistical and an econometric perspective. I believe that the paper is the richer for this. Anant Balakrishnan wrote a program to fit the parameters of the negative binomial distribution and condensed negative binomial distribution, removing a task for which I have no fondness. It was well appreciated.

Financial support from the Commonwealth Fund of New York through a Harkness Fellowship and Telecom Australia, the Australian telecommunications authority, made this paper possible and is gratefully acknowledged.

Finally, I owe sincere thanks to my wife, Pamela. Both her encouragement and logistical support have been most helpful in coping with the crash program involved with the study. She was able to shoulder her troubles, my troubles, and still have time and energy for outside activities.

## 1. Introduction

### 1.1. Objectives

The objective of this paper is to study the reliability with which respondents report long distance call usage and to consider the relevance of findings concerning systematic response errors to market research studies of other consumer purchase reports.

Self or household informant reports of purchasing and/or consumption levels are included in almost all marketing research surveys and generally are the measurements which play a critical role, either as the criterion variable or an explanatory variable, in the statistical analysis carried out on the data collected.

Concern about the presence of systematic and random error components in such data is as old as the marketing research itself and represents a key reason for the controversy and criticism that has been a continuing part of the field's history. While this concern has led to efforts to minimize bias through sampling methods, intervening techniques, questionnaire design and similar tools for discrimination and resolution, there have been relatively few efforts to assess the nature and magnitude of measurement error in reports of purchasing and consumption patterns directly. See, for example, the work by Sudman (1964a, 1964b), Parfitt (1967), Lipstein (1975), and Wind and Lerner (1979).

Where this has been done the results are not encouraging, indicating sizable amounts of systematic bias ("response effects"). Sudman and Bradburn (1974) give a review of the magnitude of response effects for different task variables and with different techniques for soliciting information (face-to-face interview, telephone, mail). For non-salient items or those involving social values they report substantial misreporting.

At the same time as controversy over survey errors has raged, there has been a growing body of statistical techniques developed in econometrics (e.g. Griliches (1974)) and psychometrics (Joreskog (1978)) that permits adjustment and correction for "errors in variables" or "unreliability". These methods rely on assumptions about the behavior of errors which have not undergone much empirical testing. For example the standard error model relating reported purchases (X) to actual (T), $X = T + \varepsilon$, used in statistics (Cochran (1968)), econometrics (Griliches (1974)), and psychometrics (Lord and Novick (1968), Joreskog (1978)) assumes independence of T and $\varepsilon$, a condition which Sudman and Bradburn (1974) consider unlikely to be met. Exceptions to this inattention to the model's assumptions include Murray's work on international trade (in which parallel forms were used rather than actual values) (Murray (1972)) and that of Neter (1970).

Operational procedures, while sometimes noting these assumptions in passing (e.g. Sudman and Bradburn (1974), p. 4), frequently ignore them in practice. The reason for this is quite clear. The opportunity to compare fallible measures to "true" values does not normally exist. The conduct of a survey of telephone toll call usage, together with accurate actual information from billing records provide an unusual opportunity to measure response effects and study their characteristics.

In order to study response effects, that is systematic differences between T and X, it is necessary to firstly develop a model which allows a description of the behavior of T, actual calling. Given this, the error series may be modeled. Following the study of these two distributions an analysis of the sources and magnitude of errors is presented. These are considered for different segments of the data. The implications for sampling design and the development of optimal estimators based only on reported data are then developed.

The importance of the study of response effects is important since errors from these sources are frequently of greater magnitude than statistical sampling errors (See Lipstein (1975)). Cochran (1968) shows the threat that errors on measurement pose by atternation of parameter estimates. A study of distribution of errors will ascertain whether frequently used estimators are appropriate. Systematic biases and errors are shown to lead to minimum variance estimators which suggest the pooling of data. Differences in accuracy across segments will lead to differential weighting of survey data.

Most of the above literature refers to consumer frequently purchased products, while the variable of interest here is an intangible service. The implications of this difference are discussed in Section 4.2 Differences arise because of lack of physical traces, lack of inventory effects, purchase from home rather than a retail outlet, and availability of detailed billing records.

## 1.2. Model of Reported Calls on Actual Calls

Systematic response effects, discussed in Section 1.1, were studied by relating calls reported on the mail questionnaire to actual calls obtained from billing records. This allowed a study of the distributional properties of errors and also estimation of the reliability of reporting. The model used to study the relationship of reported calls to actual was a compound one, borrowing from a number of studies. It suggests that reported calls equal actual average calls plus a bias plus an error. The error has two components; reporting error and error due to random fluctuations in monthly calls which introduce measurement error in the average actual number of calls. The most obvious alternative is a multiplicative model or some kind of intermediate model, involving Box-Cox/Box Tidwell transformations. A multiplicative model is shown to overestimate heteroskedesticity (Section 3.1), while an intermediate transformation has interpretation problems. Its development is provided in this section.

A random effects model may be used to consider the actual number of calls made by household i in period j, $t_{ij}$, in terms of the population mean calling rate, $\lambda$, an individual household effect, $(\lambda_i - \lambda)$, and a random month-to-month variation, $v_{ij}$ (e.g. Scheffe (1959), Chapter 7):

$$t_{ij} = \lambda + (\lambda_i - \lambda) + v_{ij} \qquad (1)$$

A number of distributions have been suggested for this process, the number of actual calls for a household in a given period of time. For example Ehrenberg (1959) shows that if $t_{ij}$ are Poisson given $\lambda_i$ and $\lambda_i$ is distributed Gamma throughout the population, then $t_{ij}$ is distributed as a negative binomial. Tests on the appropriateness of these distributions are performed in Section 2.

Since $\lambda_i$ is not observed, the survey estimate, $R_i$, is normally used:

$$R_i = \lambda_i + b_i + \varepsilon_i \qquad (2)$$

where $\varepsilon_i$ is a random error component in the survey estimate of the mean calling rate, $R_i$, and the true mean, $\lambda_i$, and $b_i$ is the bias. The reported may not be an unbiased estimate of the actual so it is necessary to add the bias term $b_i$, or equivalently say that $\varepsilon_i$ has mean $b_i$. Since there is only one survey estimate per household, equation (2) is clearly under-identified unless some assumption about the distribution of b across the population is made. It was assumed to be equal for all households in this analysis (i.e. $b_i = b$). An alternative assumption would be that it is linearly related to the calling rate on reported calling rate.

Survey response error was estimated as the difference between the survey response ($R_i$) and the average actual calling rate of households over 17 periods ($t_{i.}$) which from (1) is given by

$$t_{i.} = \lambda_i + 1/17 \sum_j v_{ij} \qquad (3)$$

Therefore substituting (3) into (2) the estimated model was

$$R_i = t_{i.} + b + \varepsilon_i - 1/17 \sum_j v_{ij}$$

It is worth noting that this model contains only one parameter, b and so its main purpose is to allow a decomposition of the variance into that due to month-to-month fluctuations and that due to measurement error and bias. This variance decomposition is performed in Section 3.1.2. The estimation of the two sources of variance, $\varepsilon$ and v, allows the estimation of reliability.

## 1.3 Background and Data Collection

The data used in this paper was gathered as part of a larger marketing study conducted by a major U.S. telephone company in Quad Cities, Illinois/Iowa. The objective of the study was to assess the effectiveness of different marketing strategies on residential toll call usage. As part of the measurement, households' actual usage was monitored from billing records while they also received questionnaires to obtain demographic, attitudinal, and perceived usage information.

Questionnaires were distributed in October to November 1978 to 36000 households selected from a particular geographic area which subscribed to cable television. They were preceded by an alert card and followed by two reminders (which only went to non-respondents). The second of these contained an additional set of questionnaires. Two copies of the questionnaire were sent to each household with a request that separate household heads each complete one. 15,303 households responded, a 43% response rate, and of these 5,911 returned two questionnaires. The sample consisted of both subscribers to cable television and non-subscribers. An analysis of non-response (Silk (1980)) found that cable viewers who responded did not differ significantly from non-respondents. Non-respondents in the non-cable group had significantly lower usage and in particular the class of zero callers was underrepresented. Silk suggests that this may be due the lower salience of the subject to non-users. Since the survey was by mail, no control over who the respondent was was possible. A procedure such as that adopted by Davis, Douglas, and Silk (1981) would have been desirable, at least for a subsample.

Billing data relate to 17 4-weekly periods from July 1978 to October 1979. While an advertising experiment was performed from January 1979, in a separate paper Chow (1982) suggests that the practical effects of this on calling rates may be ignored.

The actual calling data are recorded in great detail and with great accuracy making it ideal for the measurement of the reliability of reported calls. Unfortunately there are a number of shortcomings in the questionnaire (usage requested on a monthly rather than 4-weekly basis, business calls requested excluded, ambiguity of wording, and integer responses while average calling rates are not integer). The effect of these is examined in this study; they are not considered atypical of survey of this nature.

From the sample three sub-samples were selected. Firstly all of the individual actual calls of 100 households were examined. This group was selected from just one telephone exchange for the purpose of computational ease. The representativeness of this exchange is discussed in Section 4.1.1. This data base was used to study inter-purchase times in Section 2.1.

The second data base consisted of matched billing records and questionnaire data of a 20% sample of the survey response (3075 households), systematically sampled. Households returning more than one questionnaire were systematically sampled according to the overall sex ratio of respondents. For all tests in Sections 2.2 and 3.1 subsets of these responses were examined. For example in Section 2.2 only 992 households had complete call data for the 17 periods. In Section 3.1.3 a subset of 1211 is used to look at sociodemographic characteristics with only five months of actual data. Again, Section 4.1.1 examines the representativeness of these subsets.

Finally 975 pairs of husband-wife responses were matched with actual data to test the relative reliability of spouses. The households returning two questionnaires were considered not to be typical of the reliability of respondents and that was the reason why all analysis was not conducted on this group. Husband-wife reliability is considered in Section 3.2.

## 2. The Nature of Telephone Toll Call Usage

### 2.1. Importance of Studying the Process

Considerable attention is given in this paper to the underlying process which generates residential toll calls. The reasons for this twofold. Firstly observation of actual purchasing is a frequently used method of modeling consumer behavior (e.g. diary data). It is becoming increasingly important with the introduction of Universal Product Codes (UPC's) and detailed historical data on the purchasing patterns of consumers at the individual level.

Secondly, the nature of the process can be critical to the choice of methods which are used to measure it and the properties of estimates based on such measurements. Normality of individuals' purchasing behavior is the most common implicit assumption of segmentation studies, justifying least squares linear regression estimates as BLUE and used for hypothesis testing (e.g. Frank, Massy, and Wind (1972)). Common assumptions about the distribution of the Normal parameters throughout the population are that the mean has arbitrary distribution and that the variance of all consumers is equal. However weighted least squares could easily be used to incorporate different variances across the population.

Ehrenberg (1959) was the first to suggest that individuals' purchasing behavior in successive time periods could be fit by a Poisson distribution. He obtained reasonable empirical support for this assumption and further postulated a gamma distribution of individuals' mean purchasing rates (the Poisson parameter) throughout the population. Then the population distribution of the number of units purchased in a given time period is negative binomial.

This suggests exponential inter-purchase times for individual customers and more specifically, a mode of re-purchasing at time zero. Chatfield and Goodhardt (1973), finding this intuitively unappealing, compared it to an Erlang distribution (with $p = 2$), which has a mode at $1/\lambda$. When mixed with a gamma distribution it gave a distribution for purchases across the population which they called "condensed negative binomial" (CNBD). The Erlang gave marginally worse fits to their data than did the exponential.

The other variant of the Poisson assumption has been proposed by Morrison (1973). He suggests a "generalized Poisson" with mean $\lambda_i$, and variance $k\lambda_i$ for individual i, where k is a constant across the population.

Finally Robbins (Morrison and Schmittlein (1981)) departs from the gamma distribution of the Poisson parameter across the population and allows an arbitrary distribution.

The models motivated by these distributional assumptions are very different which suggests that considerable attention should be paid to the degree to which they are being met by the population. In particular it is usual to assume that the population is heterogenous with respect to location and scale parameters but homogeneous with respect to a distributional assumption. Work in this section suggests that this is not necessarily a good assumption.

Relatively more attention is given by the paper to the Poisson distribution, firstly because it is currently enjoying popularity as the most realistic for number of purchases in a time period, and secondly because much of the work on assessing model fit and reliability in this area has used it (e.g. Morrison (1973), Wildt and McCann (1980)).

Ehrenberg (1959, p. 34) suggests that the Poisson assumption is plausible as long as time intervals are long enough and he finds good fits to many frequently purchased items. The need for a long enough period stems from the lack of memory effect and mode at zero of the exponential distribution which are not considered realistic for most products. With a sufficiently long time interval the "carry-over" effect between purchases will be small. This problem is not expected to be substantial with toll calls since there is no physical inventory effect. If a substantial inventory effect (and thus "memory" in interpurchase times) were to exist, however, a month might be expected to be too short a period to ensure independence, given a median of two calls per household per month.

## 2.2. Interpurchase Times

The richness of the data includes detailed and precisely measured information about when "purchases" (the commencement of toll calls) take place. This can be used for more statistically powerful tests than just investigating whether the number of calls falling in a given period are Poisson (or condensed or generalized Poisson). Under the assumption that calls in a period are Poisson it is possible to show that inter-purchase times are Exponential (e.g. Hoel, Port, and Stone (1971a), p. 230).

Unfortunately the assumption of Normal calls in a period does not lend itself to a test of this nature. If calls are considered independent (as for the Exponential) then the probability of exactly no calls in a small time $\delta t$ would be zero (compared to the non-zero probability of the discrete Poisson). It would be possible to truncate the normal at zero and say that the probability of zero calls equals the probability of less than or equal to zero calls, but the distribution is no longer symmetric and its desirable properties are lost. This disadvantage of the Normal distribution assumption will be revisited in Section 2.3.

Four tests were used to evaluate the fit of exponential and Erlang inter-purchase times. There were:

1. The first order autocorrelation between successive purchases (calls) should be zero given either the exponential or the Erlang distribution. This was tested both on a sample of 100 customers individually, and also on the same customers pooled. Sensitivity of the pooled test to the deletion of significant individual autocorrelations was studied.

2. Under the exponential distribution, the mean should equal the standard deviation; under the Erlang (with $p = 2$) the mean should equal root

two times the standard deviation. The constancy of this relationship was examined across customers.

3. An equiprobability $\chi^2$ good of fit test was performed on the data to test both the Erlang and the exponential fits (Brieman (1973), p. 199).

4. The actual distribution of the inter-purchase times was examined relative to expected, to see if there was any systematic departure from exponential times. For example Chatfield and Goodhardt (1973, p. 828) suggest a "dead period" after a purchase, during which time consumers are unlikely to make another purchase.

2.2.1. First Order Autocorrelation between Purchase Times.

This test was adopted because in just studying the distribution of inter-purchase times, no explicit consideration is given to whether the times are temporally independent (a corollary of assuming expontial or Erlang distributions).

78 of the 100 households had more than 5 calls in the 254 days for which data was available. Of these 9 had significant autocorrelation at the 95% level; 6 positive, 3 negative. Since the autocorrelations have an approximate Normal (0, 1/n) distribution (Box and Pierce (1970)) and since the toll calls made by different households may be considered independent, a pooled $\chi^2_{78}$ test was possible to test for overall significance.

The resultant statistic of 110.53 is highly significant, suggesting a departure from independence. However, removing the nine significant households gives a $\chi^2_{69}$ statistic of 56.74, which is not significant. This suggests some heterogeneity within the population with respect to distribution as well as parameters.

Those with significant $\chi^2$ did not appear to vary in usage rates from the sample as a whole having a mean inter-arrival time of 10.7 days

(compared to that of the sample of 10.6) and a standard deviation of inter-arrival times of 11.5 days (the same as that of the sample), suggesting that firstly they would not be readily identifiable from more limited data and secondly that interrelationship of time periods is not a function of usage.

2.2.2. Relationship between the Mean and Standard Deviation.

Two estimates are available for the location parameter, $\lambda$, of the exponential distribution. The maximum likelihood estimator is the inverse of the mean of inter-purchase times, while an unbiased, efficient estimate of the Poisson parameter (which also equals $\lambda$) is the number of calls/254 days (Hoel, Port and Stone (1971b), pp. 46-47). The former uses more information from the data, but is highly unstable for customers with few calls. Thus both estimates were used. The usual sample estimate of the standard deviation was used.

Regressions were fitted to both with and without a mean. Under the null hypothesis that $\lambda^{-1} = \sigma$, the regression $\sigma = \alpha + \beta\lambda^{-1} + \varepsilon$ will give $\alpha = 0$, $\beta = 1$. When all callers (with more than 5 calls) are included, the intercept, $\alpha$, is significant (for both estimates of $\lambda$) and $\beta$ is significantly different to 1 (but not 1/ 2, the Erlang slope). However if the 29 customers making between 6 and 20 calls are excluded (the light callers), the slope is not significantly different to 1 and the intercept not significantly different to zero. Alternatively the regression through the origin has positive $R^2$ of .528 (i.e. explains more variance than the mean) and has a slope not significantly different to 1. The effect of eliminating the low callers is illustrated by the regression lines in Figure 1.

- Figure 1 about here -

Thus again the exponential assumption of mean equals standard deviation can be seen to relate relatively well to a large number of the population; but certainly not to all.

Very little can be said about the low callers (those who made five calls or less in the thirty six weeks). Chatfield and Goodhardt (1973, p. 831) make a similar point and the difficulty in accurately measuring the habits of such consumers makes generalizations across the population hazardous. Quantifying the validity threats of only studying significant users is difficult and not frequently addressed. As an extreme example Banerjee and Bhattacharyya (1976) make no apology for doing a goodness of fit test on the interpurchase times of 12 heavy users out of a sample of 300. Of course the very fact that these are low consumers does decrease their practical relevance.

2.2.3. $\chi^2$ Goodness of Fit Test for Exponential and Erlang.

58 of the 100 households had more than 15 calls during the period. This is the minimum number to allow a $\chi^2$ goodness of fit test since it yields fifteen inter-purchase times and thus the potential to make 3 cells of 5 observations. Two degrees of freedom are lost, one for the estimation of the distribution parameter and one because the sum of cell sizes must equal the number of calls made. Brieman (1973, p. 209) suggests that 5 is a practical minimum for the number of observations per cell.

The $\chi^2$ tests were done by forming cells each of 5 observations and calculating the expected number in each cell given the sample estimate of the distribution parameter. Thus high callers were able to be tested with more degrees of freedom, giving a more powerful test. The null hypothesis of an exponential distribution was rejected at the 95% level for 28 out of the 58 households.

Pooling the 30 non-significant households also gave a significant $\chi$ suggesting that they were also significantly different from an

exponential distribution. (Pooling is justified since households may be assumed to make toll calls largely independently of each other). There was a slightly higher proportion of rejection amongst higher callers but this may well be a reflection of the increased power of the test. Examining segments of calling rates (e.g. those who made between 15 and 20 calls) led to a rejection in each instance.

The Erlang distribution (with p = 2) gave a considerably worse fit. 50 out of the 58 households were rejected as Erlang. On only 4 households was a better fit than exponential obtained and then never significantly better.

Thus there is strong evidence to suggest that the exponential distribution is appropriate for a minority of households, at best, and that resorting to the Erlang distribution is not a move in the right direction. This prompted an examination of the departures of actual inter-purchase times from expected, to determine where the distributional assumptions were being breached.

2.3.4. Distribution of Deviations of Actual against Expected Times.

An examination of actual inter-purchase times shows that a greater number of times fall close to zero than would be expected under the exponential distribution. This is contrary to the "dead time" hypothesis of Chatfield and Goodhardt. It explains why the Erlang distribution gave a worse fit than the exponential.

This concentration of low inter-arrival times may be illustrated in two ways. Firstly for the 58 individual exponential $\chi^2$ goodness of fit tests, 52 had the actual number in the first cell greater than the expected number. This occured 27 out of 28 times when the test was significant and 25 out of 30 times when it was not.

As a second check, because the above test may be distorted by time of day effects, the expected number of inter-purchase times of less than

24 hours was compared to the observed for the 78 households making more than 5 calls over the period. For 57 of these households the observed number was greater than the expected number, for 12 it was equal (to within half a call), and for only 9 was it less than would be expected from the exponential distribution.

This suggests that calls may be generated by (at least) two different types of stimuli. Firstly a toll call may generate the need for another toll call (or there may be joint causation). This would lead to the bunching of calls which is evident in the data. Secondly random exogenous effects could stimulate calls which would explain the longer inter-purchase times. However to model the process is beyond the scope of this paper which is to test the calling process against commonly advanced hypotheses and to test the reliability of reporting of the process.

In summary, the study of the exponential assumption has to be rejected on the basis of the $\chi^2$ tests. However a number of the features of the exponential distribution (independence of inter-purchase times, mean and standard deviation being equal, mode at zero and monotonicity) can be accepted at least for a majority of households. A more thorough study of the process would not only test this distribution against an alternative mixed distribution, it would examine the effect of time of day, day of week, and monthly periodic patterns (using spectral analysis). The mixed distribution could consist of one component which relates to the stimulus of one call causing another (say a uniform distribution with a low mean inter-purchase time) and the exponential distribution reflecting random exogenous stimuli.

An alternative approach would be try different distributions. Banerjee and Bhattacharyya suggest the Inverse Gaussian distribution for which they report improved fits and forecasts over the exponential. Its non-zero mode makes it unlikely for this application.

## 2.3. The Number of Calls in an Interval

While inter-purchase times were used to give more powerful tests of the Poisson distribution (and the condensed Poisson), an examination of the number of calls in an interval is also necessary since the violence of the breaches found in exponential inter-purchase times may not be felt as strongly when considering the number of calls in a period.

Since there were only data for 17 4-weekly periods, although it would have been possible to do a $\chi^2$ test for the Poisson assumption (on one degree of freedom per household with three cells), it would not have been possible for the Normal distribution (with its two parameters).

Therefore the hypothesis of individual i making calls with a Poisson distribution, parameter $\lambda_i$, and $\lambda_i$ being distributed $\Gamma$ throughout the population was examined in two parts. Firstly the Poisson assumption was tested by looking at the independence of consecutive time periods and by examining the relationship between the mean and variance in addition to the findings in Section 2.2.

Secondly the $\Gamma$ assumption was tested by estimating the $\Gamma$ distribution on individuals' mean calling rates over the 17 periods and comparing them to the 17 sets of NBD parameters obtained for each period. Additionally forecasts of the number of calls made by zero callers in one period in the subsequent period was compared for the NBD, the Condensed NBD (Erlang inter-purchase times rather than exponential), and the Robbins estimator which assumes an arbitrary mixing distribution (Morrison and Schmittlein (1981)).

The fit of the Normal distribution was examined by standardizing each individual's calls over the 17 periods and concatenating them over the 972 households. Transformations were also considered, as were the implied theoretical properties of the Normal assumption.

2.3.1. Independence of Time Periods.

Under the Poisson assumption successive time periods are uncorrelated. Chatfield and Goodhardt (1973, Appendix) show that under the Condensed Poisson the correlation is negative and has an absolute maximum of .13.

16 of the 992 households did not make calls and thus autocorrelations were not calculated for them. Of the remaining 976, 75 were significant at the 95% level. A pooled test on the total 976 households gave a $\chi^2_{976}$ statistic of 1146, also leading to rejection of the hypothesis of no autocorrelation between periods. 541 of the autocorrelations were greater than zero and for the 75 significant households, the figure was 61. Thus the Condensed Poisson receives no support, relative to the Poisson in this test.

Upon deletion of the 75 significant households the $\chi^2_{901}$ statistic was 738, no longer significant. Thus only a minority of the households violate this condition. These tended to be heavy callers (with a mean of 10.54 calls per 4 weeks as opposed to the overall sample of 4.82).

2.3.2. Relationship between the Mean and Variance.

Estimating $\sigma^2 = \alpha + \beta\lambda$ for all 992 households gives an estimate of $\alpha$ which is marginally non-significant ($t = 1.93$) and an estimate of $\beta = 2.32$ ($t = 39$), with an $R^2$ of .6. This is not consistent with the condensed Poisson which suggests an asymptotic variability ratio, $\sigma^2/\lambda$ of .5, but is consistent with a generalized Poisson (Morrison (1973)). Regressions through the origin gave good fits ($R^2$ of .42 - .60) relative to the mean (i.e. the variance being independent of $\lambda$). However estimates of $\beta$ were highly sensitive to high callers. That is deletion of high callers significantly changed $\beta$. Additionally, as may be seen from

Figure 2, the spread increased as $\lambda$ increased, suggesting a multiplicative model.

While it was the low callers (with high inter-purchase times) who were adding instability to the exponential fits, here the high callers are having a destabilizing effect. This may be seen from the effect on the regression line of deleting high callers (also shown in Figure 2).

- Insert Figure 2 about here -

It may well be that these high callers merit specialized study and marketing attention in any case.

2.3.3. The Gamma Assumption.

The Gamma Distribution was fit by the method of moments applied to the mean calling rate for each household over the 17 periods, assuming that this would approximate the true mean (Hoel, Port, and Stone (1971b), p. 31). This may be compared to the NBD parameter estimates found by the method of moments in Table 1. Parameter estimates were obtained using an iterative program based on the formulas of Morrison and Schmittlein (1981, p. 1015). The NBD parmaters are relatively stable and are similar to those estimated directly for the Gamma. A graph of the distribution of mean calling rates shows a monotonic decrease from zero, suggesting the right shape for a Gamma. The CNBD parameter estimates are also reasonably stable. However the parameter estimates from using the method of fitting zeroes are markedly different for the NBD. The actual number of zero callers ranged from 45% to 63% of those that would be expected from the method of moments parameter estimates. A similar bias was noted with the CNBD.

- Table 1 about here -

Morrison and Schmittlein suggest the Robbins estimator as an alternative for prediction of period 2 calling rates given calling rates in period 1, particularly for the class of zero callers. The Robbins estimator of the number of units that will be purchased in period 2 by those who purchased n units in period 1, is the number of units purchased in period 1 by those who purchased n + 1 units. It is consistent and unbiased. Thus the zero class in period 1 is forecast to make as many calls in period 2 as those who made one call in period 1.

Not surprisingly the parameter estimates obtained from the fitting of zeroes gave better results for forecasting the future calls of the zero class than did the method of moments estimates for both the CNBD and NBD. Variances were all comparable and were comparable with the Robbins estimator (all standard deviations were between 42 and 45 calls). However the biases of the NBD and CNBD were markedly worse than Robbins. The average biases of the NBD were 92 and 172 calls (fitting zeroes and method of moments respectively). The average biases of the CNBD were greater; 136 and 186 respectively. These compared unfavorably with the average Robbins estimator bias of 79 calls.

These results suggest that the zero class is smaller than would be expected and that the movement from the zero class is much more prevalent than would be expected with an NBD or CNBD distribution.

2.3.4. The Normality Assumption

Section 2.2. pointed out that the normality assumption for number of calls in a period can be at best an approximation since calls are discrete rather than continuous and also non-negative. One method of examining whether this is a problem is to see the number of households for which $\mu < \sigma$ and $\mu < 2\sigma$. For these households 16% and 2.3% of the normal tail will be cut off respectively. In fact 32% of households satisfied the first condition and 77% the second. This was related to number of calls; being less of a problem with high callers. For example only 24 out of 96 with average calls of greater than 10 had $\mu < 2\sigma$. However 68% of the 383 callers with an average of less than 2 calls per month had a mean of less than 1 standard deviation.

This problem may be part of the cause of a strong positive skew present in the distribution of calls. To examine the empirical distribution each individual's calls were standardized to have mean zero and variance 1. The resultant calls were concatenated and are ploted as Figure 3.

- Figure 3 about here -

Pearson's measure of skew for this curve gave $S_k = .767$ (Kendall (1943), p. 81). Kendall's alternative measure of $\mu_3^2/\mu_2^3$ gave $331 \times 10^8$.

This strong skew together with evidence in 2.3.2. that the spread of variance increased as the mean increased (e.g. Figure 2) suggests that a log normal distribution may be more appropriate.

Logarithms of calls were taken, with $10^{-3}$ being inserted in place of zero, and the calls were again standardized, concatenated and plotted. This yielded negative skew values of $S_k = -.850$ and $\mu_3^2/\mu_2^3 = 3.2 \times 10^8$, indicating that logarithms are an overtransformation. This was sensitive to the small value chosen to replace zero, suggesting that the method is not particularly satisfactory.

The objective of a transformation in analyses such as these tends to be manifold. Not only is its purpose to assure that distributional assumptions are met, it also aims to guarantee additivity in models, linearity, and homoskedasticity. Therefore the use of some intermediate transformation (notably roots) will be deferred until the section on modeling.

## 2.4 Conclusions

The Poisson distribution was shown to have a number of short-comings in terms of fit, both directly and with its implied exponential inter-purcahse times. Adopting a Condensed Poisson (with Erlang inter-purchase times) compounds the problem of too many calls bunched around zero. Allowing the variance to be a constant times the mean (Morrison's Generalized Poisson) provides some relief although this constant is sensitive to the number of calls made by the household.

While the Poisson does not appear to give a good fit, some aspects of it may be able to be used. For example Wildt and McCann (1980) use the relationship between the mean and variance of the Poisson to obtain generalized least squares estimates of a segmentation regression model. There is enough support for the relationship between $\sigma^2$ and $\lambda$ here to allow that procedure. With arbitrary distribution the least squares estimates will remain BLUE, they will not however in general be BUE.

The normal distribution with its strong skew and its large truncated tail at zero, appears to have serious limitations as an underlying assumption. This must temper any hypothesis testing and model evaluation which is done with least squares regression. The breach of symmetry may not be as serious if it is accompanied by a similar breach with reported data. In most applications it is the distribution of the error, reported-actual, which is more relevant than the distribution of either variable.

## 3. Reliability of Reporting

### 3.1. Bias and Distribution of the Error

The 992 households in Section 2.3 all returned questionnaires, although only 485 included total information concerning the calls made by the respondent, by other household head, and by other household members.

Respondents reported "typical number of monthly long distance calls over 50 miles". Billing records of actual calls relate to lunar months (four-weekly periods) so the data was analyzed both with and without an adjustment for this potential difference in period.

The model used was that developed in Section 1.2:

$$t_{ij} = \lambda + (R_i - b - \lambda) - \varepsilon_i + v_{ij}$$

where $t_{ij}$ is the number of calls reported by household i in period j, $R_i$ is the survey response of household i, b the bias, and $\varepsilon_i$ and $v_{ij}$ the reporting error and period-to-period variation respectively.

#### 3.1.1. Relation of $t_{i.}$ to $R_i$.

Initially the calls of each household averaged over the 17 periods were examined. It was assumed that $\Sigma\, v_{ij}/17$ would be small relative to $\varepsilon_i$ and the subsequent error partition performed in this section confirms this.

While a cross-tabulation between $t_{i.}$ and $R_i$ showed a strong level of association (for example 84% of reported values were within $2\sigma$ of the average and 65% were within one standard deviation) the correlation coefficient was only .354. In fact the deletion of 13 households for whom calls were reported greater than or equal to 30 or actual average calls were greater than or equal to 30, increases the correlation to .639. The deletion of 8 households for whom reported calls were greater than or equal to 30 increases $\rho$ to .546. A scattergram plot of actual calls against reported is given as Figure 4. Morrison (1972) points out the effect of

measuring discrete variables as inputs to continuous models. Since reported calls were uniformly integer this leads to a possible maximum "error" of a half a call even with perfect respondent information. To test the effect of this, min (.5, $t_{i.} - R_i$) was added or subtracted from the reported. The correlation between this adjusted reported and actual was increased only marginally, from .354 to .370.

- Figure 4 about here -

This low correlation was thought to be due to ambiguous wording of the question in that an examination of a number of responses suggested that some respondents had confused "total reported calls for other household members" with "total calls for the household". However eliminating all households for whom this could have been the case led to a deterioration in the correlation.

Given all of the observations, a regression of actual on reported was run. The regression equation may be re-expressed to show the optimal estimate of a household's mean calling rate, given the population mean and its reported value:

Thus $t_{i.} = 2.78 + .364\ R_i + \varepsilon_i'$ $\quad R^2 = .125$

$\qquad\qquad (6.42) \quad (8.32)$ (t - statistics in brackets)

becomes $t_{i.} = p\ \lambda + (1 - p)\ (R_i - \text{bias})$ where p is the weighting constant

$$= .634 \times 4.82 + .364\ (R_i - .787)$$

Thus the optimal forecast is 63.4% of the population mean + 36.4% of the reported after adjustment for bias.

A table showing average number of calls made, reported calls, the reporting error ($t_{i.} - R_i$), and the optimal error from the above regression are contained in Table 2, together with standard deviations.

- Table 2 about here -

The figures in brackets are after adjustment for the difference between the calendar month reported and the 4-week month billing records. It may be seen that this accounts for almost half of the bias, but very little of the variance.

What is slightly disturbing to note is that if the population mean is used as an estimate, it has a lower variance in predicting a household's mean calling rate than does the survey (i.e. $\sigma^2(t_{i.}) < \sigma^2(t_{i.} - R_i)$). Thus the individual difference effect ($R_i - \lambda$) appears less than the reporting error effect, $\varepsilon_i$.

- Figure 5 about here -

The distribution of the errors, shown in Figure 5 is postively skewed but less so that $t_{i.}$. ($S_k$ = .00266 and $\mu_3^2/\mu_2^3$ = .0001 compared to .00371 and .0667 respectively). However an examination of the marginal density of the errors with either $t_{i.}$ or $R_i$ shows heteroskedacticity (Figure 6).

- Figure 6 about here -

In fact the correlation between the absolute errors and $t_{i.}$ is .638. This suggests the need for a transformation and is consistent with Section 2.3 in that respect. A Box-Cox transformation on $t_{i.}$ and Box-Tidwell transformation on $R_i$ (e.g. see Weisberg (1980), pp.137-141) is one way of overcoming that problem. The difficulty of such an approach is that after the transformations it is very difficult to attach meaning to the resultant errors and their distributions. Since that is the main thrust of this paper that approach was not pursued.

3.1.2. A Variance Decomposition

In this process, it has been pointed out that differences between actual purchases in a given period and reported purchasing rates stem from two sources; period-to-period variations and error (and bias) in reporting. This distinction is well drawn by Morrison (1973). Period-to-period variation represents variability in the true score, while error in reporting plus period-to-period variation determines variability in the observed score. Thus, if it is possible to partition the error sum of squares into month-to-month variation and reported variation from mean purchasing rate, a direct measure of reliability is available.

An analysis of variance type model is the most common way to perform the partition of the variance. The model outlined at the beginning of Section 3.1 could be considered a random effects model (and methods of Scheffe (1959, Chapter 7) adopted), except as Wildt and McCann (1973) point out, the "period effects", $v_{ij}$, are unobservable. Additionally the random effects model would need to be extended to include errors in measurement variation in the $\lambda_i$. Morrison (1973) proposes overcoming the problem by his generalized Poisson assumption already discussed. Implicitly he also makes the analysis of variance assumption that cov $(\varepsilon, v) = 0$. That is, the two sources of error are uncorrelated, an assumption to which we will return.

Wildt (1976) points out that for regression models the cov $(\varepsilon,v) = 0$ assumption is not consistent and proposes an alternative. It should be noted that the estimate of $\lambda_i$ $(R_i)$ is not derived from a regression model so his invalidation does not necessarily apply to our data. His alternative proposal, however, is clearly invalid since it assumes cov $(R_i, \varepsilon_i) = 0$, a condition which has been shown not to hold. Wildt and McCann (1980) in developing Wildt's model into a generalized least squares regression retain the assumption of the independence of $\lambda_i$ and $\varepsilon_i$.

While cov ($R_i$, $\varepsilon_i$) is unambiguously non-zero, it is very difficult to answer the question about the independence of $\varepsilon_i$ and $v_{ij}$. The classical analysis of variance approach is to partition the sum of squares as follows:

$$\Sigma_j\Sigma_i(t_{ij} - R_i)^2 = \Sigma_j\Sigma_i(t_{ij} - t_{i.})^2 + 17\ \Sigma_i\ (t_{i.} - R_i)^2$$

where $t_{i.}$ acts as an estimator for $\lambda_i$ and thus the cross product term of the errors is zero by construction. This approach was in fact adopted for the original data, for the data after bias correction, and for logged data. The analysis was also repeated on the data with 12 extreme data points deleted. For logs, all values were increased by .5 calls to overcome the problem of zero values. The results are included as Table 3.

- Table 3 about here -

It may be seen that the reliability (ratio of the variances) for original data is only about .20, but on deletion of 12 outliers and adjustment for bias, discussed in Section 4.1, it is raised to .41. These figures are consistent with the square of the correlation coeffecients between reported and actual in Section 3.1.1 ($\rho$ ($R_i$, $t_{i.}$) = .354, or .639 after deletion of outliers; $.354^2$ = .13, $.639^2$ = .41).

Young and Young (1975) point out that although deletion of observations and adjustment of zero values to allow the taking of logs are both unpalatable, with the former introducing validity problems and the latter bias, deletion of observations is likely to be preferable from a statistical point of view. Their analysis is limited to zeroes in the dependent variable and they only consider adding a constant to zero observations, not to all observations. They also only consider one constant (unity). However the analysis here was repeated for the 173 households who did not report zero calls and who made non-zero calls in each month. The results were

similar with reliabilities being somewhat lower; .153 and .263 for original data and logged data respectively.

The benefit gained from the logged data bears closer examination. The logarithmic transformation does two things; it allows a non-linear model to be fit to the data, thus potentially increasing the amount of explanation that the one variable can give for the other (which is desirable). The other effect though, is to "squash" the large data points (and inflate the very small ones) which may give a spurious increase in covariation if calculated on the transformed data. To investigate the effect of this, only respondents who reported non-zero calls and who made non-zero calls over the period were examined (405 out of 485). The correlation on original data was .315. The correlation on logged data was .600, an impressive increase. However, if the fitted data is retransformed and $R^2$ and thus $\rho$ recalculated on original data, it reduced to .360! Therefore it appears as though most of the improvement is coming from the non-linear transformation discounting those observations with the worst fit.

In conclusion the assumption of $\rho(\varepsilon, v) = 0$ deserves revisiting. It is difficult to test this assumption since our construction of the error series $v_{ij}$ ensures that the sample correlation would be zero. It would be possible to hold out some of the periods, estimating $t_{i.}$ from only a portion and then find the correlation estimate of $r(\varepsilon, v)$ on the holdout sample. The disadvantage of that is that if the samples were then switched a different result would be expected to occur. One indication of $\rho(\varepsilon, v)$ which can be gained is that the sample correlation between the standard deviation of $v_{ij}$'s (for each i) and the absolute sample reporting errors is .348. This relationship between the distribution parameters does not necessarily imply a relationship between the variables, but it is circumstantial evidence.

However in view of the lack of firm evidence invalidating the variance decomposition performed, the lack of a satisfactory alternative, and the uncertain properties of transformed data (e.g. Box-Cox, Box-Tidwell transformations) this measure of reliability is considered appropriate.

### 3.1.3. Socioeconomic Differences

Segmentation by socioeconomic characteristics has two potential uses. Firstly it may be used as an alternative source of explanation for different calling rates instead of reported calls. This may be performed using regressions and correlation coefficients. For categorical independent variables the use of linear regressions corresponds to an analysis of variance between categories.

Secondly, differences in call variances and reliability of reporting between socioeconomic groups may lead to stratification of surveys, varying weight of segments when calculating optimal estimates of calling based on reported calls, and differential marketing strategies. An alternative method of measuring differences between segments is to examine each separately. To do this a larger sample was employed (containing 1211 housholds). These households subsume the sample in the previous section.

Socioeconomic variables considered were sex, workforce status, education, age, income, and household size. The first four related to the respondent, the latter two to the household. Education had ordinal scaling; age, income, and household size ratio scaling. $R^2$ using these as explanatory variables in simple regression on $t_i$. (now based on a j of 5 periods) ranged from .002 (for household size) to .017 (education). However because of the large sample size all regressions were significant. Including all of the socioeconomic variables into a single regression did not introduce substantial multicollinearity (condition number = 14), but it only improved the $R^2$ to .033, not comparable to that of reported calls (.175 for this data set).

These low $R^2$s are not unusual for socioeconomic variables in segmentation studies. They led Frank (1967) and Twedt (1964) to conclude that market segmentation based on socioeconomic measurements was impossible. Morrison (1973) attempted to explain these low $R^2$s in terms of month-to-month variability from the individual's mean purchasing rate. However as Wildt (1976) points out, Morrison's results with their implicit assumption that cov $(v_{ij}, \varepsilon_i) = 0$, cannot apply to regression models. Beckwith and Sasieni (1976) give a formula for the upper bound which the variability of individuals' purchasing behavior imposes on $R^2$ for a perfectly fitting regression model (their Formula (23)). They suggest in most instances that this will be close to 1, and certainly greater than .5. Therefore they conclude that low $R^2$s such as those experienced here and those of Frank, Massy, and Wind (1972) are indeed due to misspecification (possibly due to omitted variables) or a breach of the generalized Poisson assumption. The latter has already been found to be suspect in this study.

Bass, Tigert, and Lonsdale (1968) point out that generally it is group means in which a marketing manager is interested rather than individual purchasing patterns. For segmentation data used by Frank, Massy, and Boyd with median $R^2$ of less than .10 they use a conditional probability model to predict group purchasing behavior and get useful discrimination. They conclude; "The inability of socioeconomic variables to explain a substantial part of the variance of usage rates does not necessarily imply that there are not substantial differences in the mean usage rates for different socioeconomic segments. Differences in mean usage rates among segments is sufficient condition for the development of a strategy of market segmentation." Wildt and McCann (1980) point out that if observations are available on multiple time periods (as is the case here)

then the temporal variation around a mean purchasing rate may be modelled directly using generalized least squares. In this case the the model would be very much like the one already advanced at the beginning of this section, with $Z_i\beta$ replacing $R_i$:

$$t_{ij} = Z_i\beta + \varepsilon_i + v_{ij}$$

This can be estimated using an Aitken estimator (Dhrymes (1970), p. 150). However as Wildt and McCann (1980, p. 338) show, the $R^2$ adjusted for within household variation (their $R^2_\lambda$) equals the group means $R^2$ estimate. Since the analysis in this section was done on group means ($t_{i.}$ rather than on $t_{ij}$) there is no further room for improvement of fit and the only use of GLS would have been as an additional method of performing the variance partition in 3.1.2.

Instead it was decided to examine correlations and variances within the segments, as opposed to differences in means which have been examined up until now. While the differences in means between segments (measured by the simple regressions above ) were small (but significant), the difference in reliability between segments was quite marked in some instances. Reliability was measured as the correlation between $t_{i.}$ and $R_i$, bias, and error standard deviation. It is reported in Table 4

- Table 4 about here -

A marked pattern emerges. Lower socioeconomic groups tend to have a higher reliability than higher ones. This may be partially due to the exclusion of business calls (discussed in Section 4.1) and partially due to a lower mean calling rate, but neither of these two factors seems sufficient to explain all of the differences.

Household size does not appear to be particularly important, neither does age. However, again due to the large sample size, these differences are all significant (the confidence limits are non-overlapping). The difference between men and women is in favor of women but is not marked and this is discussed further in the next section in which the responses of husbands and wives from the same household are examined.

In addition to socioeconomic status, a stratification was performed by actual usage. Reliability for different levels of actual usage was similar for all callers with a mean calling rate of under 20 toll calls a month ($\rho$ = .1 to .2). It was considerably lower for those making more than 20 a month ($\rho$ = .04), perhaps reflecting in part the possible exclusion of business calls and in part the greater variance in calling rate.

In each segment correlations between $t_{i.}$ and $R_i$ are lower than for the sample as a whole. This is not surprising since by taking a subsample, the mean (of $t_{i.}$ and $R_i$) becomes a better approximation to their value so that $t_{i.} = \mu$ is competing more keenly with the alternative hypothesis, expressed in regression form, $t_{i.} = a + bR_i + e_i$.

Since all reliabilities so far have dealt with all telephone subscribers regardless of whether they made toll calls or not, an estimate was also made for users of the toll service in the five month period. It could be argued that this would be a better indication of users of the toll service, presumably the market of interest. 104 of the 1211 sample members were excluded and the correlation dropped from .419 to .397. Indeed when the period of study is extended to 17 months the proportion of non users drops from 8% to 2%, suggesting that the majority are not "hard-core" non-users.

3.1.4. Accuracy of Reporting for Other Houshold Members

Since actual data on calls by respondent, other household head, and other household members were not available separately from billing records, the influence of other household members' calls on the reliability of the respondents' reporting was examined in three ways. Unfortunately each has limitations.

Firstly correlation between $t_{i.}$ and calls reported by self ($R1_i$), calls reported made by other household head ($R2_i$) and other household members ($R3_i$) were examined to see if $R1_i$ was better correlated with $t_{i.}$ than $R_i$ (= $R1_i + R2_i + R3_i$). It might not be surprising to find such an increase in reliability if responses of the same individual were compared to actual usage over time. However it was not surprising that any relationship between calls reported made by respondent and actual total calls was not consistent across the population. That is, all respondents did not make the same proportion of the household's calls (to within a constant).

Households of size one were examined. In these the repondent may generally be assumed to have made all of the calls (and this was checked to be the case). However the other distinguishing characteristics of households of size one relative to all households make this measurement suspect. The correlation did not improve for households of size one.

The third approach was to examine $R_i$ separately for those households in which the respondent claimed to have made all of the calls. A very obvious bias of "unaware respondents" and unconscientious respondents prejudices this test. Again no improvement was gained.

Thus, in summary, little can be said about the reliability of reporting for other household members habits and the fact that each of these is a worse measure of actual calls (each with a correlation of .32) than their total, suggests that these intercorrelations would not have been a good surrogate measure for reliability, had the actual not been present. Some additional evidence is presented in Section 3.2.2 in which it was possible to compare husbands' and wives' responses.

Mulaik (1972, p. 66) gives a possible theoretical explanation for the lower correlations between the components of a total with another variable than between the total and that variable. He shows that if there is a low inter-component correlation (and here they range from .26 to .46) then totalling them will improve the correlation with the external variable.

## 3.2. Husband-Wife Reliability

### 3.2.1. Relation of Actual Means to Reported

Since a number of households submitted two questionnaires as described in Section 1.2, a comparison of the accuracy of wives' and husbands' responses was possible. This allows an unusual opportunity to test the value of multiple-respondents as a method of increasing reliability. While the response is potentially biased the results are very strong and they deserve some attention. In the 396 husband-wife households that were randomly selected with total historical call data and responses to the usage questions, the results which were obtained are included as Table 5. Results for the sample in Section 3.1 are included for validity comparison purposes.

- Table 5 about here -

An examination of the data showed two husband response outliers; reported total calls and calls by self of 97 and 98. These two increased the variance of husband response error responses by 74% and decreased the

correlation between husband and wife responses from .848 to .556 and between actual and husband responses from .662 to .434. Their deletion changed the optimal weighting of wife/husband responses from .96/.04 to .74/.26, bearing witness to the sensitivity of these measures to outliers.

Since blank boxes on the questionnaire were coded as 99 and since both 97 and 98 could be easily mistranscriptions of 99 (the next largest reported figure was 44) these cases were deleted from further analysis.

It may be noted that not only is the variance of wives' responses less than that of their husbands, the variance of the error of their responses is less and their average bias is lower, also reflected by the significantly higher correlation between wives' responses and the actual than that of their husbands. Correction for the 4-weekly measurement of actual calls reduces wives' and husbands' biases to .12 and .38 calls per month respectively and their error standard deviations to 3.93 and 4.42. It does not change any of the correlations to 4 significant figures. (It would only be expected to change those involving the error term.)

A cross tabulation between husband and wife response shows that 71% of responses are within one call of each other in reported total calls and 51% are equal. This raises the possibility of collaboration which, unlike a similar study by Davis, Douglas, and Silk (1981), was not controlled. However as with the off-diagonal elements of the $t_{i.}$, $R_i$ cross-tabulation on Section 3.1, large differences were sufficient to give a correlation of husband-wife response of only .848. The correlation between the husband and wife's response errors was .741. Husband's error was highly correlated to the difference in responses, while wife's was not so strong (correlations of .506 and -.204 respectively). This leads to a restatement of the previous evidence: when the wife and the husband disagree, the husband is likely to be wrong!

Indeed this can be formalized. Granger and Newbold (1977, p. 270) have done considerable work on the combination of forecasts but their theory applies to any estimates at all. They show that the minimum variance mix of two estimates, $e_1$ and $e_2$, say $ke_1 + (1 - k)e_2$ is given by

$$k = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_1^2 - 2\rho\sigma_1\sigma_2}$$

Applying that to husband and wife responses leads to an optimal mix, R, of

R = .74 Wife + .26 Husband.

Using just wives as respondents (if this subsample were to be representative) leads to more weight being placed on the survey response, derived in Section 3.1. The optimal estimate of $t_{i.}$, given wife's response $R_{wi}$ and the population mean, $\lambda$, now becomes

$$t_{i.} = 1.22 + .603\ R_{wi} \qquad R^2 = .428$$

$$(5.5)\quad (19.7) \qquad \text{(t-statistics)}$$

$$\text{or } t_{i.} = .397 \times 3.96 + .603\ (R_{wi} - .584)$$

where 3.96 is the population mean, .584 the bias (slightly different from before because households with incomplete husband information were added), and (.397, .603) the weights given to the population mean and the survey response respectively.

The Granger and Newbold optimal husband-wife mix gives even more weight to the survey response; 31% to the population mean and 61% to the optimal mix of the two respondents, explaining 51% of the variation from the mean (i.e. a correlation between the mix and actual calls of .715).

A summary of the standard deviation of errors under the alternative methods of estimating the household's usage is given in Table 6.

- Table 6 about here -

A study of the distribution of the reporting errors showed that both were reasonably symmetric and similar in shape (although obviously not scale). However it was interesting to note that the error of wives increased markedly in magnitude as the number of calls increased ($\rho$ (actual, absolute error) = .607), while that of husbands was not so strong ($\rho$ (actual, absolute error) = .270).

As was found in Section 3.1 both wives' and husbands' absolute errors increased with reported usage (correlations of approximately .6), again suggesting heteroskedasticity and the need for a transformation.

3.2.2. Socioeconomic Differences

The socioeconomic analysis performed in Section 3.1.3 was able to be repeated with the corresponding husband-wife samples. Differences in the socioeconomic variables related to calling patterns were again studied using a correlation matrix. With the exception of reported bill, other variables including reported income, education, and household size give small but significant correlations with both actual and reported calls (all below .2 for husband and wife). The reported bill correlated well with both reported calls (.659 for wives and .617 for husbands) and actual calls (.609 and .613). It is interesting that the wife is a better informant when calls are considered, while the husband is a better informant when the bill is considered. These correlations also suggest that the Granger and Newbold technique of combining the responses for bill and calls would lead to a lower variance in reporting error. The use to which this information was put, however, was as an indication of reliability, outlined in Section 2.3.4.

The difference in husband and wife responses for employment status deserves special attention because of the different mixes between sexes. Retired respondents were found to be the best group for both sexes, consistent with their status in Section 3.1.3 (with correlation for husband, $\rho_H$, = .844, for wife, $\rho_W$ = .873). Homemakers also were above average ($\rho_W$ = .825). The low correlation for full-time employment (of respondent) in Section 3.1 can be seen to be largely a result of the wives' responses. Husbands who are full-time employees give about average accuracy reports ($\rho_H$ = .651 compared to total sample of .662), while wives who are full-time employees give a correlation of $\rho_W$ = .599 (compared to an overall $\rho_W$ of .705). The situation is even more extreme with part-time work where wives give a correlation of .488. A study of error variances gave similar results. Students and Unemployed were not examined separately because of small cell sizes but in a grouped test continued to give very high accuracy responses.

The trend of increasing education leading to lower reliability (with the exception of those having undertaken post-graduate work) which was found in Section 3.1 has two components. For both sexes those with post-graduate work are excellent respondents ($\rho_H$ = .811, $\rho_W$ = .836). For husbands the correlation decreases monotonically with increasing education up until the end of college (.91 for finished grade school, .70, for finished high school, and .57 for finished college). For wives it moves in the opposite direction (.60, .70, and .71 respectively). Looking at error variances (removing the effect of different standard deviations of actual and reported calls and a linear relation between reported and actual) led to both spouses having decreasing reliability with education, but the husband more so.

The findings of Section 3.1 with respect to age were again repeated, for both spouses. Middle aged respondents (26-40) gave the most reliable responses ($\rho_H$ = .785, $\rho_W$ = .766) with both the young and the old being lower. This contrasts interestingly with employment where retired persons gave the most accurate responses, but an examination of the data show these persons to be a small minority of the 40+ age group. Error variances showed monotonic decreasing reliability with age.

Again the relationship of increasing income being accompanied by decreasing reliability was found for both spouses. For those earning less than $10,000 the correlations were $\rho_H$ = .880 and $\rho_W$ = .785, while for those households on over $10,000 the correlations were .655 and .699 respectively. Similar results using error variances were found for wives, but for husbands the results were reversed.

3.2.3. Accuracy of Reporting for Other Household Members

As pointed out in Section 3.1.4 it is impossible to obtain actual calls for individuals within a household. However with this sample it is possible to compare the calls the husband reported that he made relative to the calls reported for other household head by the wife and conversely. It is also possible to see the amount of agreement on calls made by other household members.

Correlation between reported calls for the husband (husband's report and wife's report) is .78 while that for the wife is .86. Calls made by other household members have a correlation at .709 between husband's and wife's responses. This correlation is helped by many joint responses of zero.

Perhaps more interesting is the differences in means and the distribution of the difference. Wives and husbands gave similar average results for calls made by the husband (wives reported 1.05, husbands 1.19), for the wife (wives reported 2.19, husbands 2.34), and for other household members (wives reported 1.01, husbands .99). Both spouses agree that on an average, wives make about twice as many toll calls as husbands. Standard deviations are also similar, with the husbands being marginally higher in each case. In fact an examination of the differences in these responses shows symmetry with the exception of a small tail caused by a few husbands reporting markedly higher calls than wives. For example the distribution of the difference in wives' and husbands' perceptions of other household members' calls is given in Figure 7.

- Figure 7 about here -

One potential use of this data of two respondents is to use a composite decision rule not only for the mix of husband and wife's responses in

determining an optimal household estimate, but also to use the separate components which each respondent advances. For example, the hypothesis that each respondent most reliably reports his or her own calls was tested by using calls reported by each household head for themselves plus an average of calls by others in the household as a measure of the reported household usage. This did in fact marginally increase the correlation with actual from .705 for wives and .602 for husbands to .716 when both reported their own calls.

The converse hypothesis (that the respondent is not objective in reporting his or her own calls and is thus biased) was also examined looking at a measure of total calls made of those reported for the other household head plus an average for other household members. With a correlation with actual of .676 this was in between husbands and wives in terms of reliability.

It is, perhaps, interesting to note in passing that the correlations between husbands' and wives' reporting of household members' calls satisfies the conditions for multi-trait multi-method discriminant and convergent validity suggested by Campbell and Fiske (1959) since all measures of the same members have significantly higher correlations than within respondent or with other members. This is reflective of the relatively high level of agreement between respondents.

3.2.4. Husband-Wife Differences as Measures of Reliability

Response error is obviously not normally directly observable but if differences in respondents' answers can be found correlated to error then this may be incorporated into the estimation process by differential weighting of observations or the adjustment of responses for predicted error.

Thus this Section briefly examines whether multiple respondents could be used not only to increase reliability, but also as an indication of any individual household's reliability and the direction and magnitude of their probable error. Four measures of husband-wife differences in reported calls were available; the difference in reported calls made by husband, the difference in reported calls made by wife, the difference in reported calls made by other household members and the total difference in reports. The correlation of these four measures with husband's error in reporting and wife's error in reporting is given in Table 7.

- Table 7 about here -

Correlations with wife's error are around .2 for most of the measures and she is the more important respondent since she is more reliable. Correlation between husband error and the difference in total calls between the two spouses of .506 is to some extent an indication of the greater accuracy of the wife. Thus some gains could be made as a result of these correlations but they would not be dramatic.

Differences in husbands' and wives' responses to socioeconomic questions were also used as an indicator of reliability. Income and household size are the only two which could be used because they refer to the household rather than the respondent. Correlations between husband and wife response differences for these variables and the error of call reporting was in the order of .05; not an incisive predictor.

While husband-wife differences showed limited promise as a reliability measure, it was decided to relate the difference between the reported bill size to the reported number of calls for each respondent using an average cost per call over the whole sample. In practice this gross variable could in fact be refined. For other products it would be more clearly defined. The correlation between the difference in reported bill size and reported calls gave a correlation with reporting error of .616 for husbands and .500 for wives, despite the fact that bill size was measured on grouped data. Thus we have established an observable indicator for reliability which is available on an individual household basis.

## 4. Validity Considerations

### 4.1. Internal Validity

Using the framework proposed by Campbell and Stanley (1963) validity is considered under the headings of internal and external validity. Internal validity refers to the validity with which we are able to make inferences on relationships from our study, while external validity relates to the validity with which findings can be generalized across situations, for example across products and populations.

As a major grouping under internal validity Cook and Campbell (1979) include statistical conclusion validity. A number of threats have been posed to that in this study. Some, such as the effect of measuring mean calling rates as integers rather than continuous were able to be bounded. Others have just been raised without being evaluated (for example the seasonal and time of day effects on actual calls). Yet others, not considered central to this paper, have not been pursued at all. (For example after deletion of the significant autocorrelations in Section 2.1.1, it is unclear that the $\chi^2$ test is still appropriate, the statistic now being more in the nature of an order-statistic.)

The most obvious statistical challenge to the validity of results is the serious breach of both the Poisson and Normal assumptions in Section 2. It is the reason why little formal testing was performed in Section 3 on reliability and some restraint was placed on the use of the word "significance" in that section.

Threats to construct validity are usefully considered as threats arising from content problems and threats arising from construct limitations by Bohrnstedt (1970). Content of the questionnaire has already been criticized in a number of areas; the difference in period with that used for

actual data, the exclusion of business calls, and the wording of "total calls for other household members" which is positioned on the questionnaire where a total for the whole household might be expected to be placed. The first and third of these were specifically tested and found to not present a problem. The second presents some difficulty but the strength of reliability measure obtained shows that it is not overwhelming.

The major potential construct threat is the wording of "typical month" to probe the mean calling rate for each household. Again the evidence suggests that this was able to be done by respondents. However Ehrenberg (1959, p. 41) suggests a method of wording questions to elicit usage rates that minimize this type of bias. Under the negative binomial distribution probing usage in two periods of different length (both recent) allows the estimation of NBD parameters directly and minimizes memory problems.

Since estimates of the NDB parameters are available from the actual data, such a question in this case would have provided an interesting cross-validation. In cases where the actual data are not available such a technique is necessary to estimate a household's NBD parameters.

The other two apparent threats to internal validity are those of diffusion of treatments and that of selection. The mailing of the two questionnaires and lack of control over who filled them in suggests that it may have been performed jointly or even by one spouse in many instances. There are no data with which to test that hypothesis.

The potential problems with selection biases are discussed in the next sub-section.

4.1.1. Selection Biases

The survey was a geographic census of the area in which it was conducted so no bias as such exists in the respondent selection in the survey as a whole. The choice of area and its representativeness are discussed under the heading of external validity.

As pointed out in Section 1.2 an analysis of non-response with respect to billing data was able to be performed; yet another unusual aspect of the data set. Some bias toward high users was evident in responses. Unfortunately an early half-late half analysis of responses was not performed to judge the likely reliabilities of non-respondents.

However some data were available for the population as a whole which received questionnaires. The overall calling rate was 4.8 calls per month which is the same as in Section 3.1. The variance (calculated from pooled data) falls somewhere between 7 and 9 calls per month which includes the sample standard deviation of 8.4. An examination of the distribution of mean calling rates shows that it is similar to the sample used for Sections 3.1 and 2.2. Thus non-respondents seem to have similar calling rates to respondents although their reliability and socioeconomic characteristics may be different.

From the total respondents three sub-samples were selected as indicated in Section 1; one of 100 households to examine detailed calling records, one of 3075 households to examine calls in a period and also for the basic reliability work, and one of 975 households for which both husband and wife returned questionnaires.

The first data base was only used to examine the distribution of time between calls. Households had a mean number of calls of 3.9 per month which is a little lower than the population average. No socioeconomic data were available for these households.

The second group of households had a mean and variance equal to that of the sample as a whole and a similar distribution. It can be considered representative from a usage point of view. Respondents in the data base who did not complete the questionnaire or for whom there was not total historical data (e.g. they moved) were examined separately. Their composition in terms of sex, age, education, income and household size was extremely similar to that of respondents with the marginal exception that women who completed the questionnaire and for whom there was total data were four years younger than average (45 as compared to 49) and in a slightly higher income bracket (by about $1500 in $18000). However they tended to be lower toll callers with actual average calls of 2.5.

The third group of households, that of husbands and wives, is interesting to compare not only to the population as a whole, but also to the second data base to see if there are significant biases in those who returned two questionnaires relative to the average response. (The second data base also included a proportionate number of 2-respondent households).

Actual calls were a little less in the husband-wife data base than the main reliability data base and standard deviations were considerably less (the means were 3.98 and 4.92 calls per month respectively, and the standard deviations 5.19 and 8.38). Reliability was considerably greater with both husbands and wives compared to the responses as a whole. The reliability of wives of .705 and husbands of .662 compared to the overall of .354. The standard deviations of the error were correspondingly lower (4.1 and 4.65 compared to 9.40). This led to a thorough case analysis of the second data base, since the deletion of two respondents had given such dramatic results in Section 3.2.

One respondent had reported calls of 97, which again may be considered a transcription error. After deletion the correlation becomes .414. In all, eight households had reported calls of 30 or over and these respondents had an average absolute error of 54 calls per month. Their deletion on the basis of the fact that they are identifiable from the survey and their responses can be used as a screening mechanism, gives a correlation between reported and actual of .546 (males .530, females .550). On the grounds that the questionnaire wording can be accused of encouraging inaccurate responses from high actual (business) users the additional deletion of three respondents with actual calls of over 30 per month was undertaken, giving a correlationof .639 (males .598, females .662). These figures are not quite equal to those in Section 3.2, with the slightly lower correlations being expected from households which did not necessarily submit two questionnaires. They show the sensitivity of results to extreme observations. The deletion of 2 out of 396 in the third questionnaire and 12 out of 485 in the second dramatically alters the results. Only outliers which can be identified from their primary data have been deleted.

The socioeconomic characteristics of both husbands and wives were remarkably similar to those of men and women in the second data base suggesting that two responses from a household is not related to socioeconomic differences. Again the socioeconomic characteristics of those who provided complete data and those for whom complete billing records existed were similar to those of respondents as a whole.

## 4.2. External Validity

External validity considerations can be considered to fall largely into two groups; validity across geographic and social groups and validity across other products and services.

The bias introduced by considering only those persons with both television sets and telephones is not considered to be substantial. The area of Quad Cities in the Mid-West is one of a number considered in some ways "typical" of this country. While it may be a good blend of differences between the East and West Coasts and between large cities and small rural areas, clearly examining one area does not allow one to extrapolate across the country with confidence. Other considerations such as mobility of the population would be important in considering the representativeness of Quad Cities as a microcosm of America.

Extrapolating across products raises a number of characteristics of the toll service not common to many products and even some services. These impinge on both the study of the process and also reliability of reporting. Firstly it is unclear as to whether there is a natural satiation level with toll calls (as opposed to foodstuffs for example and many services such as haircuts). The marginal utility of toll calls may not necessarily decrease with increased usage. There are no inventory effects in common with most services. Finally the point of purchase is the home, making it more accessible than many products. These differences are highly likely to be the major cause why in contrast to the Erlang distribution of Chatfield and Goodhardt, we found a larger than expected mode close to zero.

With respect to reliability of reporting, the telephone bill details purchasing and thus provides a reference in a manner which would not be duplicated for many products. Thus reliability in this study might be expected to be higher. Conversely usage is only visible by direct observation or an examination of the bill (compared to, say, gas in the car or the amount of coffee in the house). The salience of the telephone ensured a good response rate and to achieve similar results for other products it might be necessary to use personal interviews or conduct the survey by telephone.

# 5. Conclusions

## 5.1. Summary

An examination of the process which generated toll calls showed that inter-purchase times were not Exponential or Erlang ($p = 2$), thus prejudicing use of the Poisson distribution in modeling the number of purchases in a period. In addition, when mixed with the Gamma assumption of households' mean calling rates throughout the population, while the result is stable over time, it did not match the highly unstable Robbins estimator in predicting the future behavior of those who made zero calls in a period. However, a number of the features of the Poisson (and particularly the Generalized Poisson) do hold. The mean is significantly linearly related to the variance. Consecutive time periods' purchases are independent for most households. And inter-purchase times are monotonic decreasing. Therefore the cost of using the Poisson for Generalized Least Squares, for example may not be particularly great.

The Normal assumption has two disturbing features. Firstly given that the number of calls in a period is typically small (with a mode of 2), the assumption of continuity is not realistic. Secondly the inherent non-negativity of calls in a period cuts off substantial tails of the distribution for most households, given the large standard deviation relative to the mean. Logarithmic transformation is too strong and intermediate transformations must be discussed in the context of the complete model (their effect on additivity, heteroskedasticity, and fit). In that context the meaning of such a model, particularly when it is the property of the error terms which is of interest, is unclear.

Reliability proceeded from that unhappy stepping stone. The "raw" reported score was found to give a worse estimate (in squared error terms) than the overall population mean. This was found to be able to be improved by the deletion of 12 out of 485 observations, 9 selected on the basis of observable data, 3 justified by the wording of the questionnaire which excluded business calls. The optimal estimate is obviously still a weighted average of the population mean and the reported score (in fact this is a degenerate case of Granger and Newbold's combination of forecasts). An analysis of variance measure of reliability gave much the same results as the correlation measure between actual and reported calls.

While meaningful relationships cannot be found relating calls to socioeconomic variables (to explain differences in means), marked differences in the covariances between segments was found (i.e. the reliability).

A study of equivalent husband-wife pairs showed the wives to be more reliable and an optimal weighting of the two responses was 74% wife, 26% husband. This was not uniform across segments and in particular working wives made extremely poor respondents. The reliability of response for wives and husbands moved in opposite directions with increasing education.

Differences between husbands' and wives' responses to the questionnaire did not give a strong indication of either partner's reliability. However a coarse measure of the level of agreement between amount of the bill and the number of calls made gave an excellent indication.

## 5.2. Areas for Future Research

### 5.2.1. Compound Estimation Rules

A number of observable indicators have been gained to measure reliability of an individual respondent (correspondence of bill size to number of calls, socioeconomic characteristics, reported number of calls, etc.). This allows a mixed optimal weighting between the population mean and the reported score to be employed to decrease the variance of the error. Alternatively the bill size could be used for such respondents or additional effort could be diverted to measuring their habits or those of a sub-sample (similar to the sampling of non-response).

The relative reliability of husbands and wives varies across socioeconomic segments and this could be incorporated into the estimation process. Work done by Granger and Newbold in relation to the combination of forecasts seems well suited to this task. The important point to note is that the best answer is not to accept the respondent's response, without adjustment. Clearly if it is group means that are of interest rather than individuals (and Bass, Tigert, and Lonsdale (1968) suggest that this is the case), then the importance of this method is reduced. However to the extent which there is error in the group mean, benefits will still be realized.

### 5.2.2. Alternative Measures of Association

The correlation coefficient has been used almost exclusively to measure association between reported and actual calls. Being based on the ratio of cross products and sums of squares it is extremely sensitive to extreme observations and a more robust measure would provide an interesting contrast.

In addition the correlation coefficient measures the strength of a linear relationship with intercept and general slope between two variables.

If the optimal weighting of the sample mean and the survey response outlined in the paper are undertaken then that is appropriate. However if the intercept is to be constrained to zero and the slope to 1 (as would be the case if the respondents' raw answer were to be accepted), then the correlation will overstate the relationship.

5.2.3. Modeling of the Calling Process

Considerable attention was devoted to what the calling process was not. It would be interesting to model it to account for the large number of calls close together and to see if a two-stimulus model would fit the data. Seasonal effects and trends also deserve further attention.

Tracing through the implications of such a distribution on the reliability measurement and the formation of confidence levels (by simulation if necessary) would form an interesting study.

5.2.4. Reliability Measurement

The possibility of a non-linear transform, while theoretically unappealing, seems to be called for by the data (for example the skewed distribution of standardized calls and the increasing absolute error with increasing calling rates). Such a transformation must fulfill a number of objectives which may be conflicting, so it may not be possible. However the investigation of whether a Box-Cox/Box-Tidwell transformation could increase the reliability is a logical extension of this paper.

5.2.5. Modeling of Homogeneous Segments

Evidence was presented in Section 2 that population of households appeared to differ not only in terms of location and scale, but more fundamentally in terms of the underlying distribution.

Efforts directed to fitting the distribution of the inter-purchase times and the number of calls in a period, suggested in Section 2.3.4, should look seperately at different groupings within the population. Two such types of groups are relevant. Firstly groupings that may be determined on the basis of observable variables (the questionnaire responses) should be examined. Normally actual purchasing data are not available and thus it is relevant to see how well the data can be fit without them.

Secondly in an attempt to better understand the statistical process (or processes) which are generating the data, groupings which best fit the actual data should be considered.

5.2.6. Validity Considerations

It is unrealistic to expect to be able to change flaws in the method by which the data were collected. Especially since generally it appears to have been done very well. Even given the data, though, a number of techniques exist to increase the internal and external validity of results. Internal validity could be increased by examining another sample and thus allowing a split halves analysis, for example. External validity could be enhanced by extracting calling rates and socioeconomic data for other areas of the country and to ascertain whether they match those of the sample.

| | | NDB | | | | | | CNDB | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Method of Moments | | | Fitting Zeroes | | | Method of Moments | | Fitting Zeroes | |
| Period | $t_{\cdot j}$ | $\sigma_{\cdot j}$ | $\hat{\alpha}$ | $\hat{r}$ | $\hat{P_o}$ | $\hat{\alpha}$ | $\hat{r}$ | $\hat{\alpha}$ | $\hat{r}$ | $\hat{\alpha}$ | $\hat{r}$ |
| 1 | 4.52 | 8.85 | .061 | .277 | .254 | .148 | .669 | .028 | .253 | .062 | .560 |
| 2 | 4.49 | 9.96 | .047 | .213 | .257 | .147 | .660 | .023 | .207 | .062 | .557 |
| 3 | 4.42 | 8.79 | .061 | .268 | .248 | .158 | .698 | .028 | .248 | .066 | .583 |
| 4 | 4.78 | 8.84 | .065 | .311 | .239 | .144 | .688 | .031 | .296 | .061 | .583 |
| 5 | 5.28 | 8.53 | .078 | .413 | .183 | .164 | .866 | .037 | .391 | .068 | .718 |
| 6 | 5.07 | 10.12 | .052 | .264 | .246 | .126 | .639 | .025 | .254 | .054 | .548 |
| 7 | 4.69 | 8.23 | .074 | .349 | .250 | .141 | .661 | .035 | .328 | .060 | .563 |
| 8 | 4.80 | 9.45 | .057 | .274 | .245 | .138 | .662 | .027 | .259 | .058 | .557 |
| 9 | 4.94 | 10.37 | .048 | .238 | .246 | .132 | .652 | .023 | .227 | .056 | .553 |
| 10 | 5.06 | 9.60 | .058 | .293 | .212 | .150 | .759 | .028 | .283 | .063 | .638 |
| 11 | 4.56 | 9.36 | .055 | .251 | .256 | .144 | .657 | .026 | .237 | .061 | .556 |
| 12 | 4.64 | 7.86 | .081 | .377 | .219 | .168 | .780 | .038 | .353 | .070 | .650 |
| 13 | 4.66 | 7.49 | .090 | .421 | .194 | .195 | .909 | .043 | .401 | .078 | .727 |
| 14 | 4.71 | 8.22 | .075 | .354 | .211 | .172 | .810 | .036 | .339 | .071 | .669 |
| 15 | 4.77 | 10.90 | .042 | .200 | .235 | .148 | .706 | .020 | .191 | .062 | .591 |
| 16 | 4.57 | 9.59 | .052 | .240 | .244 | .152 | .695 | .025 | .229 | .064 | .585 |
| 17 | 4.72 | 9.32 | .058 | .272 | .220 | .163 | .769 | .027 | .255 | .068 | .642 |
| r | 4.81 | 8.38 | .0685 | .329 | | | | | | | |

$P_o$: Proportion of Zero Callers in the Period

$\alpha$, r: Negative Binomial, CNBD, and Gamma parameters

$t_{\cdot j}$: Average Calls in the Period (across households)

$\sigma_{\cdot j}$: Standard Deviation of Calls in the Period (across households)

TABLE 1: Parameter Estimates of the Negative Binomial Distribution, the Condensed Negative Binomial Distribution, and the Gamma Distribution using the Methods of Moments, and of Fitting Zeroes.

| | Mean | Standard Deviation |
|---|---|---|
| Average Number of Calls Made | 4.82 | 8.38 |
| Reported Number of Calls Made | 5.60<br>(5.17) | 8.15<br>(7.53) |
| Reporting Error | -.787<br>(-.356) | 9.40<br>(9.07) |
| Error of Optimal Estimator | 0 | 7.85<br>(7.85) |

Figures in Brackets represent adjustment for 4-weekly, monthly differences.

TABLE 2: Mean and Standard Deviations of Average Calls ($t_{i.}$), Reported Calls ($R_i$), Error ($t_{i.} - R_i$), and error from optimal estimator ($.64\ \mu t + .36\ (R_i - \text{bias})$).

| Source: | Original Data | | Logged Data | | Original with Outlier Deletion | |
|---|---|---|---|---|---|---|
| | Including Bias | Adjusted for Bias | Including Bias | Adjusted for Bias | Including Bias | Adjusted for Bias |
| Period-to-period $\Sigma\Sigma(t_{ij} - \bar{t}_{i.})^2$ | 187,992 | 187,992 | 3,752 | 3,752 | 84,094 | 84,094 |
| Reporting error $17\,\Sigma(t_{i.} - R_i)^2$ | 733,355 | 728,249 | 6,306 | 5,449 | 124,498 | 121,202 |
| Total $\Sigma\Sigma(t_{ij} - R_i)^2$ | 921,347 | 916,241 | 10,058 | 9,301 | 208,592 | 205,296 |
| Reliability $\frac{\text{Period-to-Period SS}}{\text{Total SS}}$ | .204 | .205 | .373 | .403 | .403 | .410 |

TABLE 3: Variance Decomposition and Estimation of Reliability of Calling Reports.

| | HUSBAND | WIFE | SECTION 3.1 |
|---|---|---|---|
| Number of Calls, Actual | | | |
| - Mean | (1.) 3.98 (2.) 3.98 | (1.) 3.98 (2.) 3.98 | (1.) 4.82 (2.) 4.14 |
| - Standard Deviation | (1.) 5.18 (2.) 5.19 | (1.) 5.18 (2.) 5.19 | (1.) 8.38 (2.) 4.30 |
| Number of Calls, Reported | | | |
| - Mean | (1.) 5.20 (2.) 4.73 | (1.) 4.44 (2.) 4.44 | (1.) 5.60 (2.) 4.78 |
| | (3.) 4.80 (4.) 4.37 | (3.) 4.10 (4.) 4.10 | (3.) 5.17 (4.) 4.41 |
| - Standard Deviation | (1.) 8.66 (2.) 5.98 | (1.) 5.37 (2.) 5.46 | (1.) 8.15 (2.) 4.78 |
| | (3.) 8.66 (4.) 5.98 | (3.) 5.37 (4.) 5.46 | (3.) 7.53 (4.) 4.41 |
| Bias of Response | (1.) 1.22 (2.) .75 | (1.) .46 (2.) .46 | (1.) .79 (2.) 1.42 |
| | (3.) .82 (4.) .39 | (3.) .12 (4.) .12 | (3.) .36 (4.) 1.05 |
| $\rho$ (Actual, Reported) | (1.) .434 (2.) .662 | (1.) .704 (2.) .705 | (1.) .354 (2.) .662 |
| | (3.) .434 (4.) .662 | (3.) .704 (4.) .705 | (3.) .354 (4.) .662 |
| $\rho$ (Error Error) (Husband Wife ) | (1.) .417 (2.) .741 | | |
| | (3.) .417 (4.) .741 | | |
| $\rho$ (Report Report) (Husband Wife ) | (1.) .556 (2.) .848 | | |
| | (3.) .556 (4.) .848 | | |

1. Raw Data
2. Deletion of outliers (2 for Husband/Wife, 12 for Section 3.1)
3. Adjustment for 4-weekly effects
4. Adjustment for 4-weekly effects and Deletion of Outliers

TABLE 5: Mean Calling Rate Together with Husband and Wife Reports.
Correlation of Husband and Wife to Actual.

| Estimator Used: | Standard Deviation of Erro |
|---|---|
| Husband's Report ($t_{i.} - R_{iH}$) | 4.65 (4.42) |
| Wife's Report ($t_{i.} - R_{iW}$) | 4.54 (3.93) |
| Population Mean ($t_{i.} - t..$) | 4.10 |
| Optimal Husband-Wife Mix ($t_{i.} - \text{Mix}(R_{iWH})$) | 3.65 |
| Using Optimal Weighting of Population Mean and Wife (Regression equation) | 3.86 |
| Using Optimal Weighting of Population Mean and Optimal Husband/Wife Mix | 3.64 |

Figures in brackets include adjustment for 4 weekly/monthly effects.

TABLE 6: Standard Deviations of Error from Different Estimators of Calls in the Husband-Wife Reliability Study.
(After Deletion of 2 Outliers)

| | Correlation With Wife's Error | With Husband's Error |
|---|---|---|
| Report of Total Households Calls: Husband-Wife Difference | -.204 | .506 |
| Report of Wife's Calls: Husband-Wife Difference | -.016 | .410 |
| Report of Husband's Calls: Husband-Wife Difference | -.208 | .233 |
| Report of Other Household Members Calls: Husband-Wife Difference | -.187 | .302 |
| Report of Bill Size: Husband-Wife Difference | .057 | .049 |
| Report of Income: Husband-Wife Difference | -.001 | -.045 |
| Report of Calls Relative to Bill Size: Husband's Difference | .326 | .616 |
| Report of Calls Relative to Bill Size: Wifes's Difference | .500 | .319 |

TABLE 7: Correlation of Observable Variables in Husband-Wife Sample with Accuracy in Reporting.

| | Segment | $\rho$ $\sigma_\varepsilon$,b | Segment | $\rho$ $\sigma_\varepsilon$,b | Segment | $\rho$ $\sigma_\varepsilon$,b | Segment | $\rho$ $\sigma_\varepsilon$,b |
|---|---|---|---|---|---|---|---|---|
| Sex | Male | .4003<br>8.2,1.1 | Female | .4254<br>6.3,.7 | | | | |
| Employment | Full-time | .3497<br>8.4,1.0 | Part-time | .2990<br>6.6,-.1 | Homemaker | .6267<br>2.8,.9 | Retired | .7382<br>5.1,1.6 |
| | Student | .8790<br>2.6,1.5 | Unemp-loyed | .8676<br>4.5,.6 | | | | |
| Education | To Grade School | .6041<br>5.1,2.4 | To High School | .5084<br>5.9,.8 | To College | .3025<br>7.9,.6 | Post-Graduate | .5808<br>8.4,1.1 |
| Age | 0-25 | .3507<br>10.0,1.4 | 26-40 | .5013<br>5.3,.9 | 41-60 | .4040<br>7.5,.2 | > 60 | .4215<br>7.2,1.8 |
| Income | $0-9999 | .8509<br>3.1,1.0 | $10000-19999 | .3730<br>6.6,.9 | $20000+ | .3838<br>7.7,.8 | | |
| Household Size | 1 | .5222<br>3.8,.8 | 2 | .4285<br>7.1,1.1 | 3 | .4337<br>7.6,.2 | 4+ | .4277<br>7.3,.5 |

TABLE 4: Correlation between Reported Calls and Actual, Standard Deviation of Errors, and bias for Various Socioeconomic Segments.
(Data Base from Section 3.1)

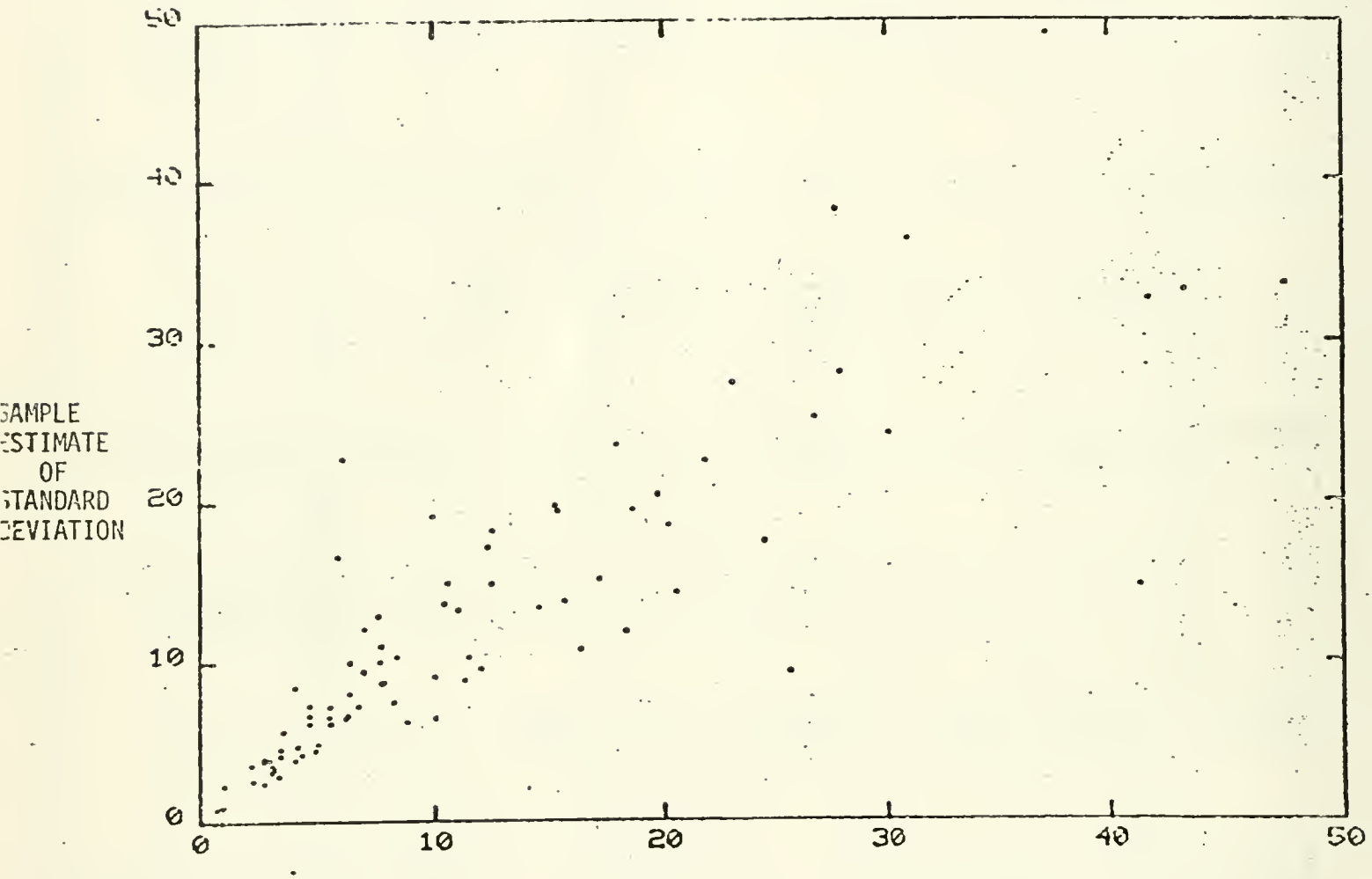


FIGURE 1: SCATTER PLOT OF MAXIMUM LIKELIHOOD ESTIMATES OF MEANS AND STANDARD DEVIATIONS FOR INDIVIDUALS INTER-PURCHASE TIMES WITH REGRESSION LINES CORRESPONDING TO THOSE WHO MADE MORE THAN 5, 12, 20, AND 32 CALLS IN THE SAMPLE PERIOD.

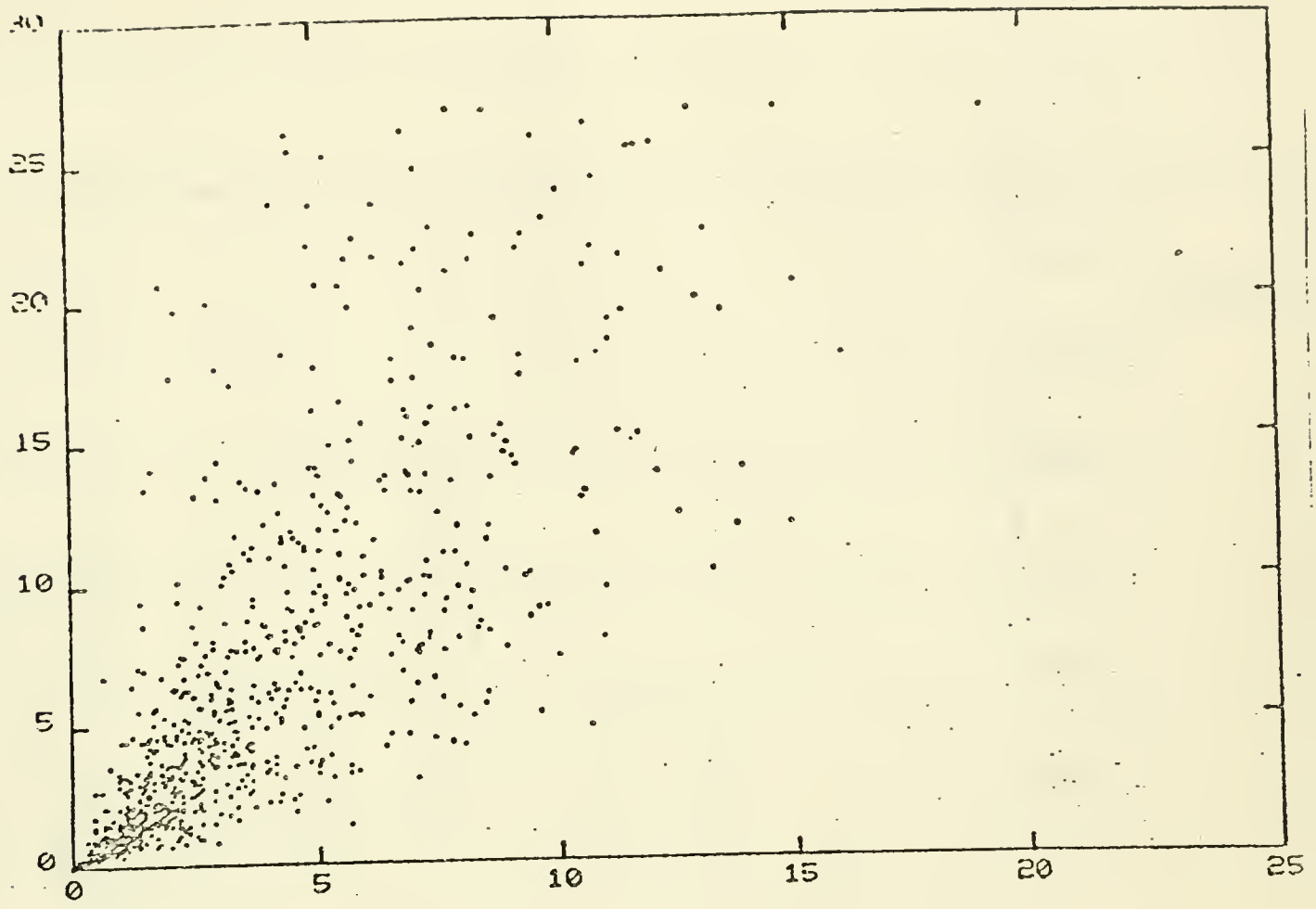


FIGURE 2: SCATTER PLOT OF ESTIMATES OF MEANS AND VARIANCES FOR INDIVIDUALS NUMBER OF CALLS IN A PERIOD WITH REGRESSION LINES THROUGH THE ORIGIN CORRESPONDING TO THOSE WHO MADE MORE THAN 1 CALL, MORE THAN 5, 12, 20 AND 32.

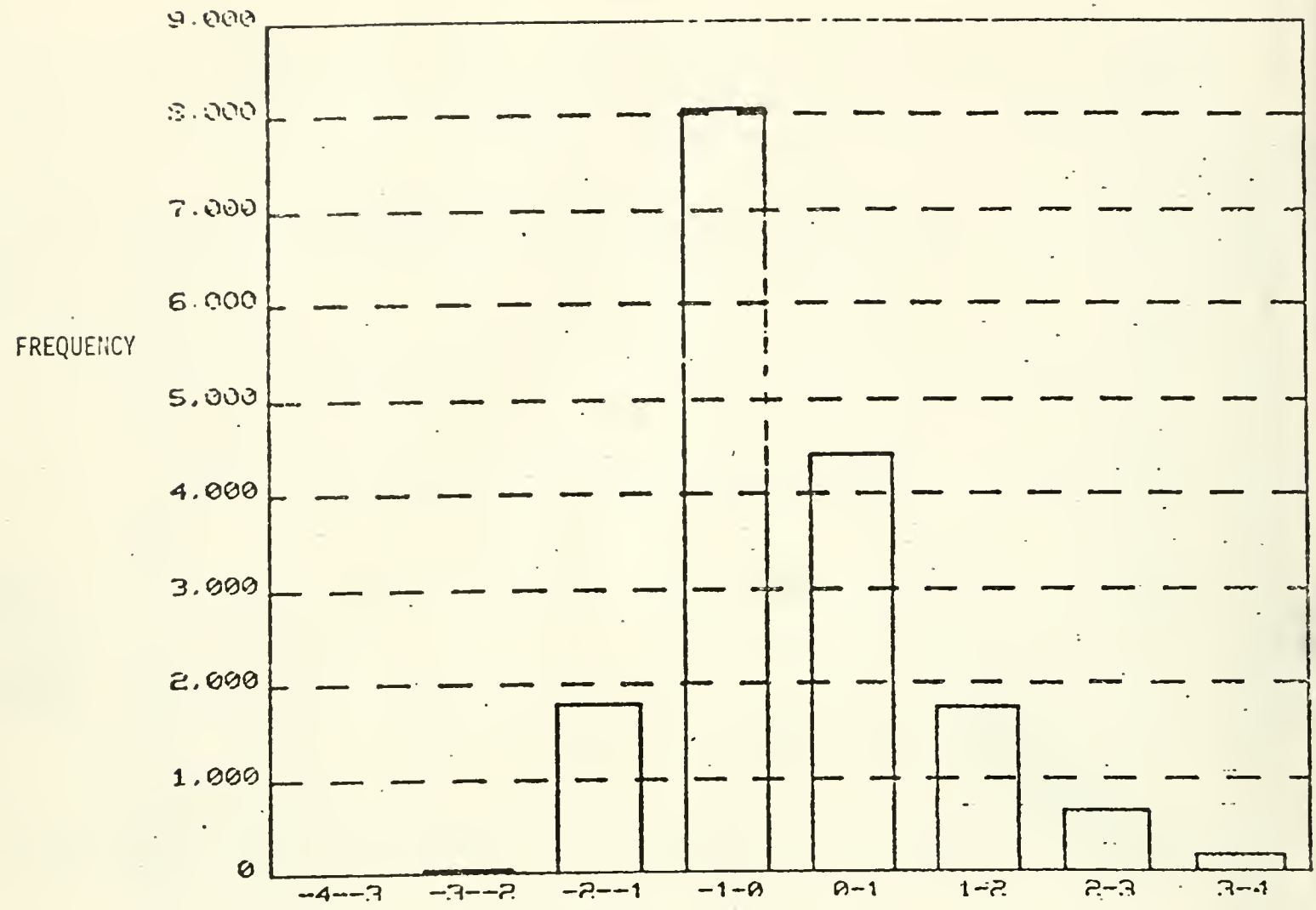


FIGURE 3: HISTOGRAM OF NUMBER OF CALLS MADE IN A PERIOD STANDARDIZED BY INDIVIDUALS' MEANS AND STANDARD DEVIATIONS.

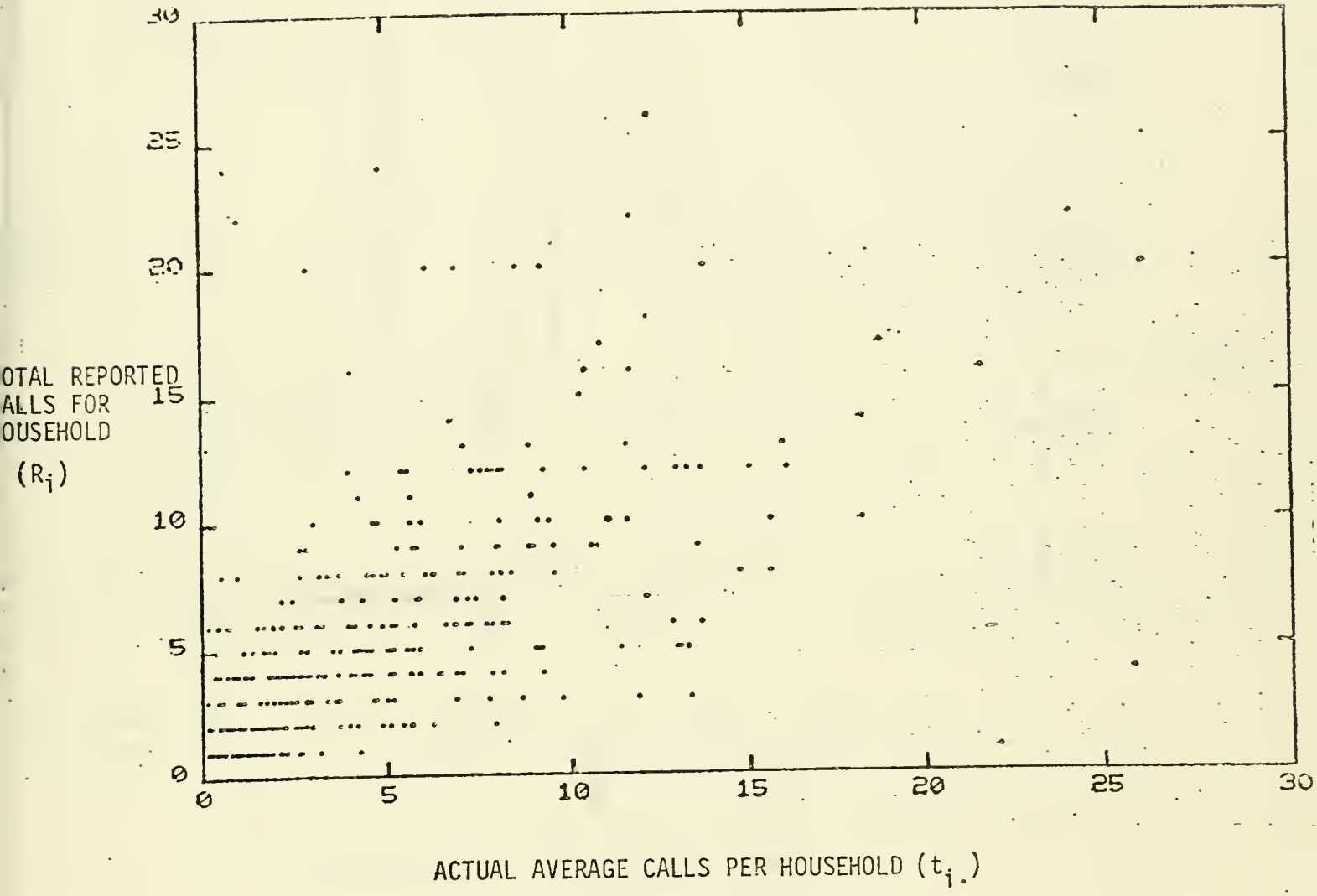


FIGURE 4: SCATTER PLOT OF REPORTED CALLS AGAINST ACTUAL CALLS FOR SECTION 3.1 DATA BASE, EXCLUDING REPORTED AND ACTUAL CALLS > 30.

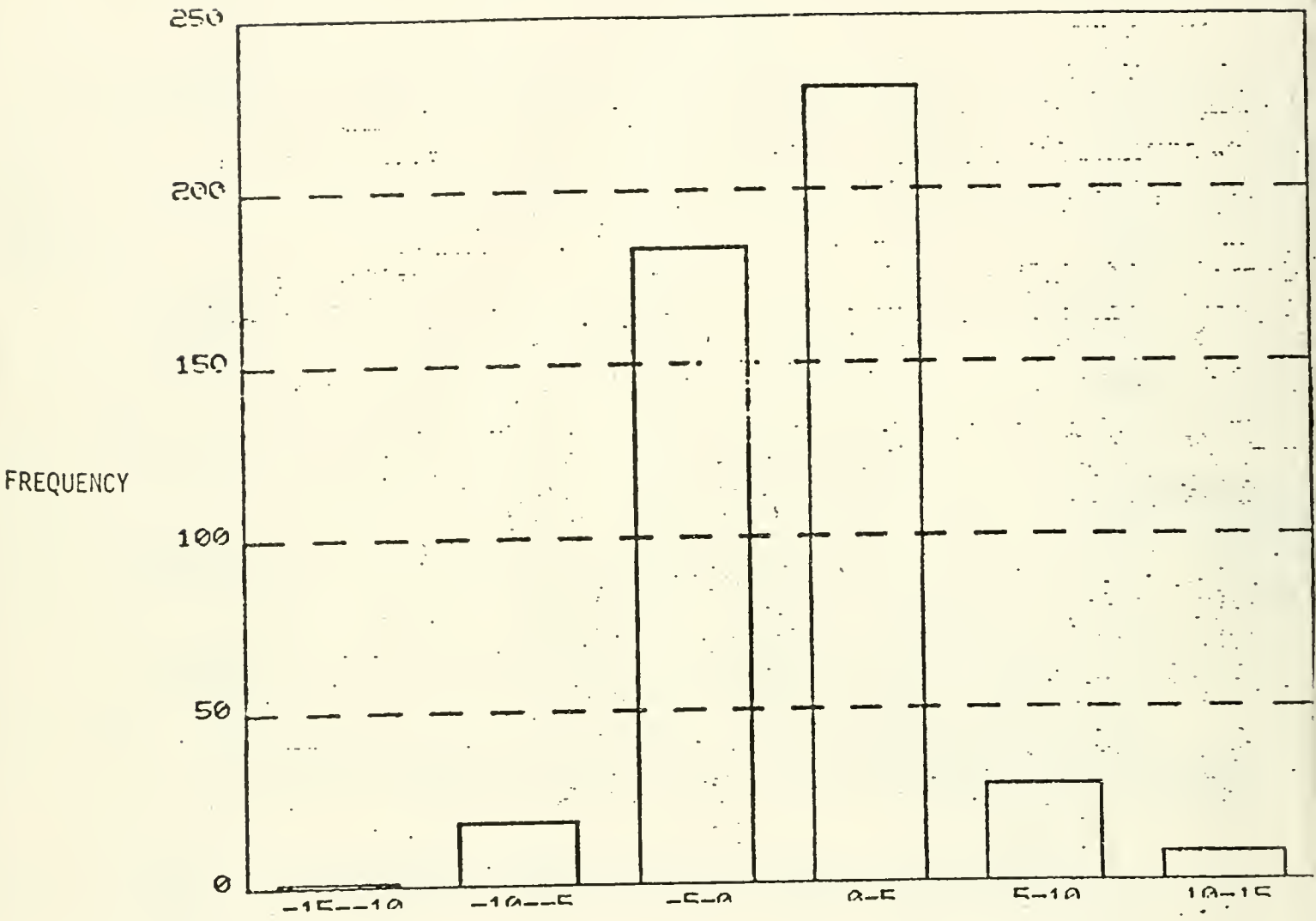


FIGURE 5: HISTOGRAM OF ERROR IN REPORTING, $\varepsilon_i$, (REPORTED CALLS - ACTUAL CALLS) FOR DATA BASE IN SECTION 3.1.

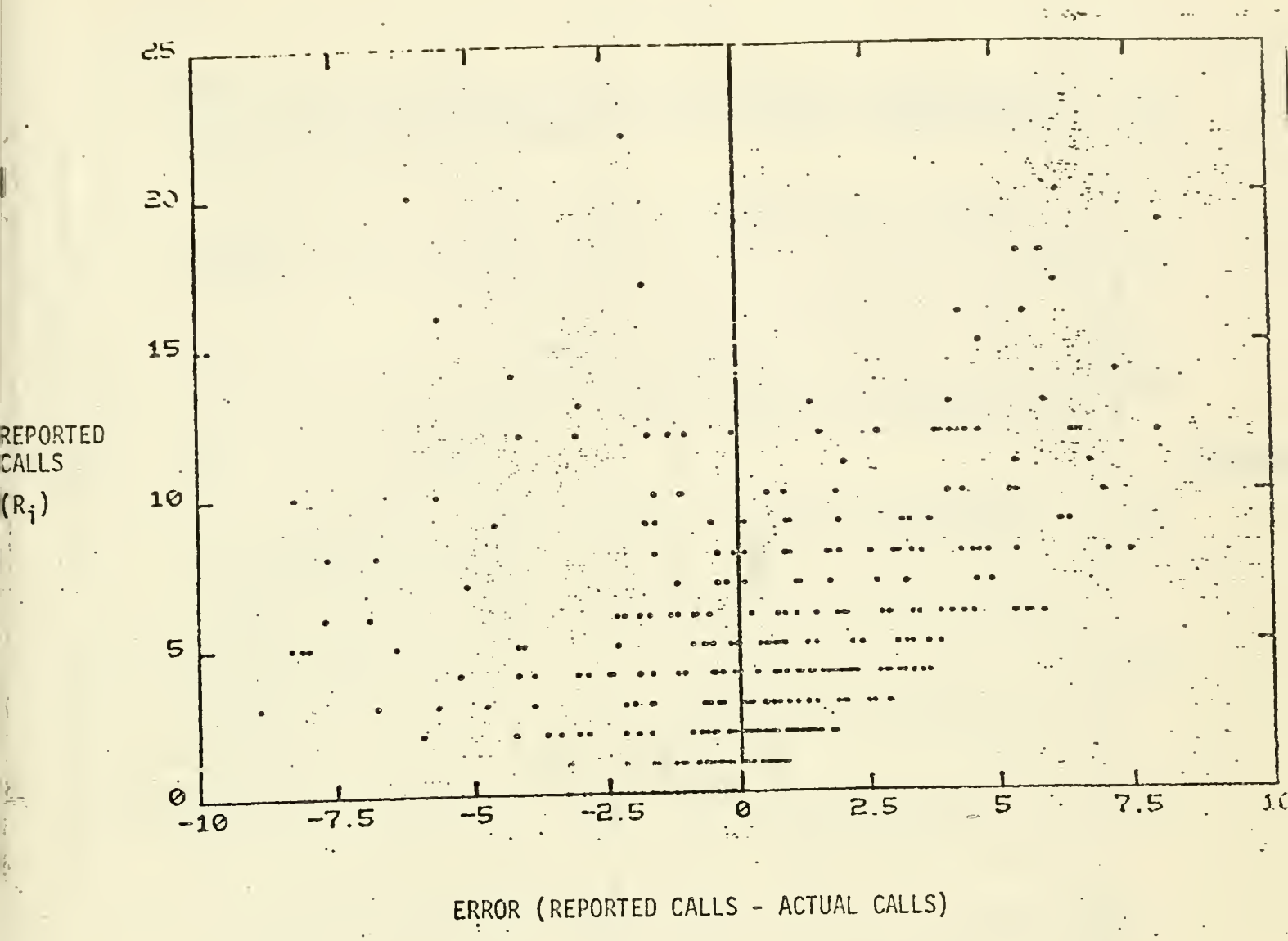


FIGURE 6: SCATTER PLOT OF ERRORS OF REPORTING AGAINST REPORTING LEVELS.

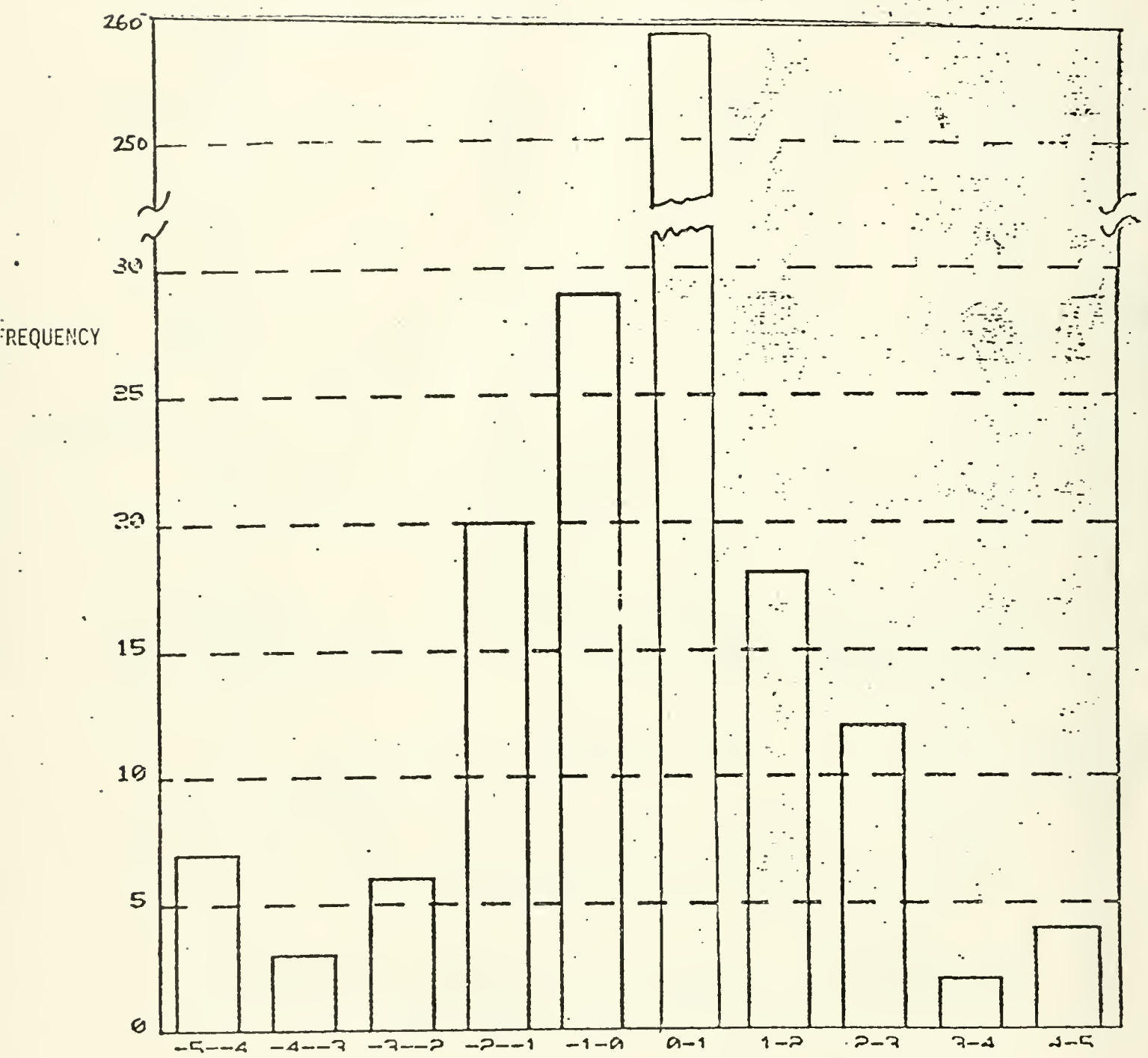


FIGURE 7: HISTOGRAM OF DISTRIBUTION OF THE DIFFERENCE BETWEEN THE WIFE'S REPORT OF OWN CALLS AND HUSBAND'S REPORT OF CALLS BY OTHER HOUSEHOLD HEAD.

BIBLIOGRAPHY


1. Banerjee, A. K. and G. K. Bhattacharyya (1976), "A Purchase Incidence Model with Inverse Gaussion Interpurchase Times," Journal of the American Statistical Association, Vol. 71, No. 356 (Applications Section, December 1976).

2. Bass, F. M., D. J. Tigert, and R. T. Lonsdale (1968), "Market Segmentation Group Versus Individual Behavior," Journal of Marketing Research, Vol. V (August 1968).

3. Beckwith, N. E. and M. W. Sasieni (1976), "Criteria for Market Segmentation Studies," Management Science, Vol. 23, No. 8 (April 1976).

4. Bohrnstedt, G. W. (1970), "Reliability and Validity Assessment in Attitude Measurement" in G. F. Summers (ed.) "Attitude Measurement" Rand-McNally, Chicago.

5. Box, G.E.P. and D. A. Pierce (1970), "Distribution of Residual Autocorrelations in ARIMA Time Series," Journal of the American Statistical Association, Vol. 65, No. 332.

6. Brieman, L. (1973), "Statistics With a View Toward Applications." Houghton Mifflin, Boston.

7. Campbell, D. T. and D. W. Fiske (1959), "Convergent and Discriminant Validation By The Multitrait-Multimethod Matrix," Psychological Bulletin, Vol. 56, No. 2 (March 1959).

8. Campbell, D. T. and J. C. Stanley (1963), "Experimental and Quasi-Experimental Deisgns for Research on Teaching" in N. L. Gage (ed.) "Handbook of Research on Teaching." Rand McNally, Chicago.

9. Chatfield, C. and A. J. Goodhardt (1973), "A Consumer Purchasing Model With Erlang Inter-Purchase Times," Journal of the American Statistical Association, Vol. 68, No. 344 (Applications Section, December 1973).

10. Chow, S. (1982), "Identifying Pre-Post Treatment Effects on Discrete Attitudinal Scales: An Application of Confirmatory Factor Analysis and Latent Structure Analysis Methods." Forthcoming, Sloan School of Management Science Working Paper.

11. Cochran, W. G. (1968), "Errors of Measurement In Statistics," Technometrics, 10 (November 1968), 637-666.

12. Cook, T. D. and D. T. Campbell (1979), "Quasi-Experimentation." Rand McNally, Chicago.

13. Davis, H. L., S. P. Douglas, and A. J. Silk (1981), "Measure Unreliability: A Hidden Threat to Cross-National Marketing Research?" Journal of Marketing, Vol. 45 (Spring 1981).

14. Dhrymes, P. J. (1970), "Econometrics." Harper International, New York.

15. Ehrenberg, A.S.C. (1959), "The Pattern of Consumer Purchases," Applied Statistics, Vol. 8 (Part 1).

16. Frank, R. E. (1967), "Market Segmentation Research: Findings and Implications" in Bass, King, and Pessemier, "Application of the Sciences in Marketing Management." John Wiley and Sons, New York.

17. Frank, R. E., W. F. Massy, and Y. Wind (1972), "Market Segmentation." Prentice Hall, Englewood Cliffs, N.H.

18. Granger, G.W.J. and P. Newbold (1973), "Forecasting Economic Times Series." Academic Press, New York.

19. Griliches, Z. (1974), "Errors In Variables and Other Unobservables," Econometrica, 42 (November 1974), 971-998.

20. Hoel, P. G., S. C. Port, and C. J. Stone (1971a), "Introduction to Probability Theory." Houghton Mifflin, Boston.

21. Hoel, P. G. and C. J. Stone (1971b), "Introduction to Statistical Theory." Houghton Mifflin, Boston.

22. Joreskog, K. G. (1978), "Structural Analysis of Covariance and Correlation Matrices," Psychometrika, 43 (December 1978), 443-478.

23. Kendall, M. G. (1943), "The Advanced Theory of Statistics," Vol. 1. Charles Giffin, London.

24. Lipstein, B. (1975), "In Defense of Small Samples," Journal of Advertising Research, 15 (February 1975), 33-40.

25. Lord, M. and M. R. Novick (1968), "Statistical Theories of Mental Test Scores." Addison-Wesley, Reading, Mass.

26. Morrison, D. G. (1972), "Regression with Discrete Random Variables: The Effect on $R^2$," Journal of Marketing Research, Vol. 9, No. 3 (August 1972).

27. Morrison, D. G. (1973), "Evaluating Market Segmentation Studies: The Properties of $R^2$," Management Science, Vol. 19, No. 11 (July 1973).

28. Mulaik, S. A. (1972), "The Foundations of Factor Analysis." McGraw Hill Book Co., New York.

29. Murray, T. W. (1972), "An Empirical Examination of the Classical Assumptions Concerning Errors In Data," Journal of the American Statistical Association, 67 (September 1972), 530-537.

30. Neter, J. (1970), "Measurement Error In Reports of Consumer Expenditures," Journal of Marketing Research, 7 (February 1970), 11-25.

31. Parfitt, J. H. (1967), "A Comparison of Purchase Recall with Diary Panel Records," Journal of Advertising Research, 15 (February 1975), 33-40.

32. Scheffe, H. (1959), "The Foundations of Factor Analysis." McGraw Hill Book Co., New York.

33. Silk, A. J. (1977), "Test-Pretest Correlations and the Reliability of Copy Testing," Journal of Marketing Research, Vol. 14, November 1977.

34. Silk, A. J. (1980), "Pre-Assessment Questionnaire Analysis, Part I: Comparison of Cable and Control Groups," mimeo, Sloan School of Management, May 1980.

35. Sudman, S. (1964), "On the Accuracy of Recording Consumer Panels: I," Journal of Marketing Research, 1 (May 1964), 14-20.

36. Sudman, S. (1964), "On the Accuracy of Recording of Consumer Panels: II," Journal of Marketing Research, 1 (August 1964), 69-83.

37. Sudman, S. and N. M. Bradburn (1974), Response Effects In Surveys: A Review and Synthesis. Chicago: Aldine, 1974.

38. Twedt, D. W. (1964), "How Important to Marketing Strategy is the Heavy User," Journal of Marketing, Vol. 28 (January 1964).

39. Weisberg, S. (1980), "Applied Linear Regression." John Wiley, New York.

40. Wildt, A. R. (1976), "On Evaluating Market Segmentation Studies and the Properties of $R^2$," Management Science, Vol. 22, No. 8 (April 1976).

41. Wildt, A. R. and J. M. McCann (1980), "A Regression Model for Market Segmentation Studies," Journal of Marketing Research, Vol. 17 (August 1980).

42. Wind, Y. and D. Lerner (1979), "On the Measurement of Purchase Data: Surveys Versus Purchase Diaries," Journal of Marketing Research, 16 (February 1979), 39-47.

43. Young, K. H. and L. Y Young (1975), "Estimation of Regressions Involving Logarithmic Transformation of Zero Values in the Dependent Variable," The American Statistician, Vol. 29, No. 3 (August 1975).

5 '82

BASEMENT

Date Due

Lib-26-67

HD28.M414 no.1303- 82
Roberts, John /The nature of residenti
744306 D*BKS 00134092
3 9080 002 013 420